UNIVERSIDAD MAYOR DE SAN ANDRES
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMATICA

TESIS DE GRADO

"RECONOCIMIENTO DE TEXTO MANUSCRITO CONTINUO EN LENGUA AYMARA"

PARA OPTAR AL TITULO DE LICENCIATURA EN INFORMÁTICA
MENCIÓN: INGENIERÍA DE SISTEMAS INFORMÁTICOS

POSTULANTE: RAMIRO BALMACEDA CHUQUIMIA

TUTOR: LIC. LUISA VELASQUEZ LOPEZ M.Sc.

REVISOR: LIC. CELIA ELENA TARQUINO PERALTA

LA PAZ – BOLIVIA

2007

DEDICATORIA

El presente trabajo no hubiera sido posible sin la colaboración desinteresada de mis queridos padres, Venturo y mi querida madre Celia por eso doy gracias a dios y dedico este trabajo a mi madre querida

AGRADECIMIENTOS

Quisiera dar las gracias a todas aquellas personas que han hecho posible con su ayuda, ejemplo y amistad, que mi camino, paso a paso, día a día, me haya llevado hasta aquí.

En primer lugar a mis directores de tesis:

Lic. Luisa Velásquez López tutora del presente trabajo, por sus conocimientos compartidos con sus alumnos en la elaboración de tesis, la forma propia de enseñanza y aprendizaje que aplica en aula, como el trabajo en grupo que considero que ha sido uno de los puntos mas importantes en compartir conocimiento e información entre alumnos, en estructuración de la tesis, metodología de investigación, operalización de variables, uso de herramientas adecuados para el desarrollo de la tesis y seguimiento continuo tomando en cuenta varios días a la semana más del tiempo asignado a la materia.

Lic. Celia Tarquino Peralta, revisora del presente trabajo, por sus conocimientos compartidos con mi persona, las observaciones a mis errores y sobre todo las sugerencias de fuentes de información para una mejor elaboración de la misma, sugerencias sobre las metodologías en pruebas de error , hipótesis y la aplicación adecuada del modelo o prototipo en los resultados obtenidos. Por todo ello gracias infinitamente Lic. Luisa Velásquez, Lic. Celia Tarquino.

Mi familia ha sido y es lo más importante para mí. Ellos me ha apoyado y guiado siempre, y soy lo que soy gracias a ellos, ni mas ni menos, ni menos ni mas. Lo primero gracias a mis padres Venturo Balmaceda Mollo a mi madrecita Celia Chuquimia Cocarico, a Grabiela Sempertegui, a mis hermanos Mario, Edgar, Cesar, Delia, Blanca, Elvira y America Balmaceda.

No quiero dejar de agradecer a uno de mis amigos Hugo Gallegos, considero uno de mis compañeros que también me apoyo en mi carrera universitaria con aquellas pequeñas cosas que hacen a la amistad.

RESUMEN

Esta tesis tiene en términos generales el estudio e implementación de un sistema de reconocimiento automático de texto manuscrito en lengua aymara, que se basa en la tecnología del reconocimiento del Habla (RH). Se hace la segmentación en el estudio de características aisladas, basándose en los "Modelos de Markov de capa oculta "(HMMs), esta tecnología se desarrolló para el reconocimiento automático de voz, debido a las similitudes con el problema del reconocimiento automático de texto manuscrito, se utilizó en este documento cuentan con mucho prestigio en este campo.

A continuación se implementa un sistema de reconocimiento de caracteres manuscritas aisladas con atributos propios del idioma aymara como por ejemplo los caracteres h' o q' que dan un significado de pronunciación diferente en aymara, el carácter apóstrofo ('), forma parte de la implementación en el reconocimiento de texto. Se realiza el reconocimiento global de las frases manuscritas para el modelado, considerando todos los atributos propios en el alfabeto del idioma aymara en el reconocimiento de patrones.

Finalmente se muestran los resultados obtenidos al implementar con los Modelos de Markov de capa oculta (HMMs), como resultado final el renacimiento del carácter y valor ASCII de cada una de los caracteres en estudio.

Sobre el problema planteado si Los Modelos de Markov de capa oculta (HMMs), Autómatas Estocásticos de estados Finitos tanto como los N-gramas son herramientas que me permiten modelar escrituras manuscritas en este caso en el idioma aymara con resultados esperados mostrados en el capitulo 4 en la prueba del modelo.

Se desarrolla satisfactoriamente sobre las bases teóricas en el reconocimiento de texto manuscrito continuo mediante los modelos estadísticos. Presenta el análisis para modelado de frases en el capitulo 2 y capitulo 3 en el idioma aymara. Se implementa el reconocimiento óptico de caracteres (OCR) en el procesamiento de la imagen off-line para el reconocimiento de texto manuscrito satisfactoriamente con resultados mucho más de lo planteado en la hipótesis.

Cumpliendo así satisfactoriamente los propósitos, objetivos específicos y los alcances planteados para la elaboración de esta tesis en reconocimiento de patrones en idioma aymara, coadyuvando así en un aporte mas en beneficio de nuestras culturas milenarias así como la escritura misma del idioma, también a partir de de un reconocimiento ya en le lenguaje del computador un usuario que no comprenda la escritura el aymara así como el significado mismo podrá traducirlo con herramientas traductores Aymara – Español por ejemplo.

CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

1 PRESENTACIÓN

1.1 INTRODUCCION

En la situación actual en nuestro país Bolivia, no se da amplia cobertura al incentivo de recuperar nuestras tradiciones culturales como son los aymaras, quechuas y otros.
Esta investigación del reconocimiento de texto manuscrito continuo en lengua aymará, no se toma en consideración, teniendo en cuenta los cambios importantes que se están dando en nuestro país con respecto a revalorizar nuestras culturas milenarias y esto supone incursionar en su contexto general de lo que puede significar una investigación en un idioma nativo como es el aymará, con sus respectivos atributos, de esta manera es que surge la propuesta de una investigación en este tema del reconocimiento automático de texto manuscrito continuo en lengua aymará, como lo indica el titulo, se hará un proceso del texto al lenguaje del computador, es por tal motivo que manifiesta importancia tanto para los que son originarios o no en este idioma.

El idioma aymará cuenta con un Grafemario Aymará de 26 consonantes y 3 vocales con sus respectivos alargamientos vocálicos, ya que en el área rural se tiene regiones en su mayoría hablantes y escritores en esta lengua, también puede incursionarse en los traductores bilingües en este idioma tan común en nuestra región.

En esta área de investigación en el reconocimiento automático de texto manuscrito continuo (escritura a mano), se puede abordar en sus diferentes ramas en la informática como también como en la estadística por ejemplo tenemos algunas aplicaciones en redes

neuronales, inteligencia artificial (IA) en el reconocimiento de formas, teoría de momentos en el área de estadística, lógica difusa y muchas otras aplicaciones que hoy en día ha marcado mucho interés en este tema en el campo científico, sus aplicaciones importantes por ejemplo tenemos en reconocimiento de cantidades numéricas en cheques bancarios ya que esto por lo general siempre se escribe a mano, reconocimiento de firmas y otras aplicaciones de suma importancia [Casacuberta V.;1987].

Esta investigación se centrará en el reconocimiento automático de escritura manuscrita en lengua aymará, ya que este idioma o lengua tiene una amplia complejidad en la escritura misma, en la forma o atributos de escritura que tiene este idioma, que es muy diferente a la escritura continuos en el idioma español, ya que el mismo presenta problemáticas como escrituras morfológicas, sintácticas y léxicas en la escritura manuscrita. Para dar una solución a estos problemas esta investigación se abordara con los modelos estadísticas siguientes: Autómatas Estocásticos de estados finitos (AEEF), "modelos de Markov de capa oculta"(HMMs) y los modelos N-gramas para el reconocimiento automático de esta escritura a mano que coadyuvarán en el desarrollo del objetivo.

En el capitulo 2 se enuncian los fundamentos, formulaciones y algoritmos con relación a los modelos de Markov de capa oculta (HMMs), autómatas estocásticos de estados finitos (AEEF) y los modelos de N-grmas.

En el capitulo 3 se implementa el reconocimiento óptico de caracteres (ORC) para el procesamiento de imágenes manuscritas al computador y conjuntamente con los modelos que se plantean.

En el capitulo 4 se realiza la descripción formal e implementación del reconocimiento de texto manuscrito continuo en lengua aymara y el análisis de datos obtenidos del modelo.

En el capitulo 5 se realizan el análisis de las conclusiones a las que se llega y las respectivas recomendaciones para futuras investigaciones.

1.2 ANTECEDENTES

Esta investigación en la actualidad esta cobrando alto interés en su investigación en el campo científico.

Tiene su aplicación en redes neuronales abordando por ejemplo en reconocimiento de patrones de caracteres, lógica difusa, agentes inteligentes, Inteligencia Artificial, Algoritmos genéticos, etc.

Los modelos de Markov de capa oculta tiene una implementación diferente mencionados anteriormente en reconocimiento de caracteres manuscritos que se vera mas adelante.
Los "modelos ocultos de Markov de capa oculta" (Hidden Markov models) HMMs esta cobrando importancia en la aplicación en este campo para su implementación y esta también manifestándose como una base de dominio para el reconocimiento de texto manuscrito continuo.

Los "Modelos de Markov de capa oculta (HMMs) " por lo general se ha abordado generalmente en la tecnología del reconocimiento del habla (RH) .

Entre los trabajos de investigación realizados en el área de reconocimiento de texto en la carrera de informática de la Universidad Mayor de san Andrés U.M.S.A. mencionamos los siguientes trabajos;

Titulo: "Reconocimiento de escritura a mano"
Postulante: Cinthya Verónica García Bellota
Objetivo: Diseñar un reconocedor de escritura a mano que incremente la precisión hasta ahora alcanzada.
Año: 1998
Titulo: "Reconocimiento de escritura a mano mediante redes neuronales"
Postulante: Teodoro Martín Aliaga Quisbert
Objetivo: Diseñar y desarrollar el modelo de un reconocedor de patrones digitales que permita reconocer un texto manuscrito o escrito a mano mediante redes neuronales.
Año: 2002
Titulo: "Sistema I.O.R(Reconocimiento de objeto invariante) para el reconocimiento de la escritura a mano"
Postulante: Zulma Heredia Poma
Objetivo: Realizar el diseño Formal de un sistema de Objeto Invariante para el reconocimiento automático de caracteres manuscritos, de tal manera que incremente la precisión del reconocimiento.
Año: 2002
Titulo: "Reconocimiento del habla mediante los modelos ocultos de markov"
Postulante: Willy Grover Catari Ramos
Objetivo: Mejorar la tasa de reconocimiento por palabra con respecto a los métodos

deterministicos, mediante los modelos ocultos de Markov (superar en al menos un 2% el porcentaje de reconocimiento alcanzado por la técnica DTW)[Glass,2003].

Año: 2006

Titulo: "Reconocimiento automático de Caracteres Manuscritos Continuos con Redes Neuronales y Lógica Difusa"

Postulante: Ramiro Alvarado Yapu Flores

Objetivo: Diseñar un prototipo de reconocimiento automático de manuscritos utilizando redes neuronales artificiales con lógica difusa, y otro utilizando solo redes neuronales artificiales, para realizar un análisis de precisión en el reconocimiento de caracteres manuscritas.

Año: 2006

La presente investigación abordará el reconocimiento de texto manuscrito continuo con los modelos estadísticos Hidden Markov models (HMMs), Autómatas Estocásticos de Estados Finitos (AEEF) para su implementación en reconocimiento de texto manuscrito en lengua aymará, los modelos de Markov han sido aplicados en reconocimiento de habla en la tesis mencionada anteriormente de la carrera de informática, el aporte para la presente investigación con respecto a los trabajos realizados en la carrera de informática con los modelos Hidden Markov models (HMMs) se implementó en el reconocimiento del habla (RH). Esta investigación en su particularidad se afrontará en el reconocimiento de texto manuscrito continuo en el idioma aymará, resolviendo problemas que mencionamos en la tabla 1.1.

1.3 SITUACION PROBLEMÁTICA

En las organizaciones o culturas tradicionales del Tahuantinsuyu que comprende todas las culturas que conocemos actualmente, que todavía mantienen algunas tradiciones por ejemplo sus idiomas, en este marco que todavía no se ha dado énfasis completa en revalorizar sus tradiciones, el originario de estas culturas especialmente en áreas rurales le es muy difícil adaptarse a otro idioma como el español, inclusive no lo pueden pronunciar muy bien el español es por eso que se sienten marginados por terceros en las instituciones académicas y otros, esta investigación podría ayudar una vez que la información este en código del computador que es la meta de este propósito el estudiante puede también

procesar la misma información traduciendo al otro idioma, utilizando una herramienta de traductor de aymará-español por ejemplo.

Tomando en cuenta problemáticas en esta área de investigación se consideran los siguientes aspectos en lo que son las problemáticas, las causas de la misma, los efectos que ocasiona y las soluciones que se planten para resolver las mismas, veamos claramente en la tabla 1.1.

Tabla 1.1 Causa – Efecto

PROBLEMA	CAUSA	EFECTO	SOLUCION
Difícil de entender en la legibilidad del documento manuscrito	Lectura incomprensible del documento	Información que no llega a procesarse	Implementar un mecanismo de reconocimiento de texto manuscrito basado en la tecnología de reconocimiento del habla (RH)
Mala escritura en los cheques bancarios	Escritura numérica poco legible	No se rembolsan el monto del dinero	Aplicar los modelos estadísticos HMMs, AEEF y los modelos de lenguaje de N-gramas
Mala estandarización en la escritura Aymará (Grafemario)	Diferentes formas de escribir en este idioma	Ambigüedad en el entendimiento de la escritura	Aplicar los modelos estadísticos HMMs, AEEF y los modelos de lenguaje de N-gramas
Reconocimiento ineficiente de Patrones en la escritura a mano por el ordenador.	No se obtiene la fuente textual adecuadamente	Información que no se procesa de manera adecuada	Desarrollar una herramienta para distinguir patrones de interés de manera textual
Estilo diferentes en los caracteres como ser: Pendiente de línea base inclinación, vertical y altura	Formas de escritura propia	Incomprensión de la frase o palabra	Implementar una herramienta de normalización de atributos de estilo
Formación incorrecta de frases morfológicas	Formas de escritura propia	Escrituras diferentes	Implementar el modelo de la maquina de estados finitos (MEF)
Formación incorrecta de frases léxicas	Formas de escritura propia	Escrituras diferentes	Implementar el modelo de la maquina de estados finitos (MEF)
Formación incorrecta de frases sintácticas	Formas de escritura propia	No existe normalización de atributos	Implementar el modelo de la maquina de estados finitos (MEF)
Escritura erróneas en la ortografía y la gramática	Formalismos de escritura individual	No existe normalización de atributos de escritura en las personas	Desarrollar un pre-procesamiento en las palabras que implemente la correcta escritura.
Migración de datos manuscritos, sin una adecuada herramienta para el procesamiento al lenguaje de maquina	Carencia de productos	Lentitud en los procesos de información	Implementar los mecanismos de reconocimiento óptico de caracteres (ORC) (como es ASCII)

Fuente: [Velásquez, Datos Propios]

1.4 FORMULACION DE PROBLEMAS

P

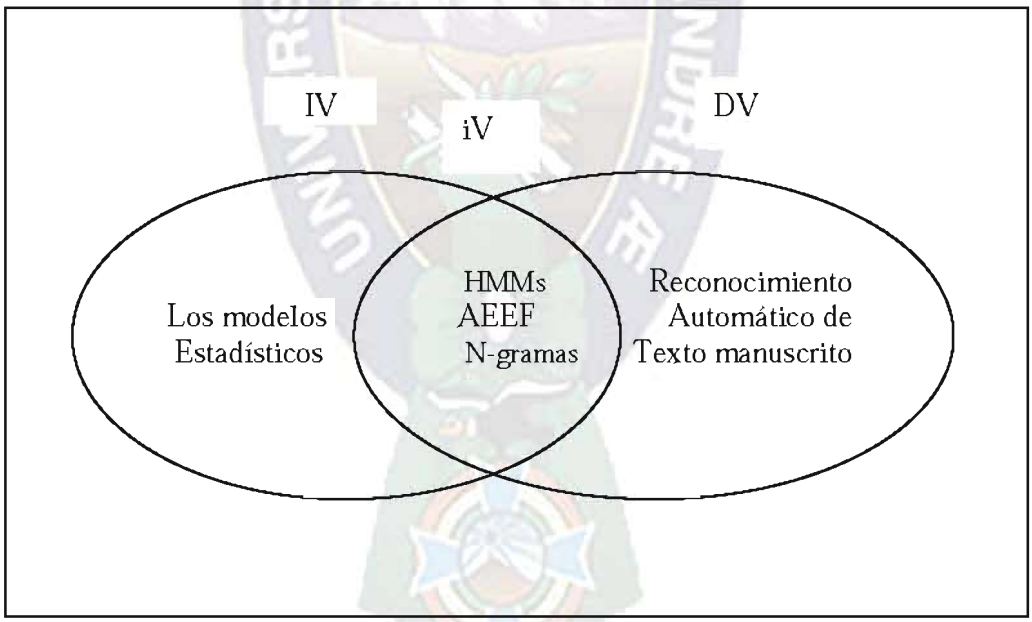
¿Los modelos de Markov de capa oculta (HMMs), Autómatas Estocásticos de Estados Finitos (AEEF) y los Lenguajes de N-gramas, será una herramienta adecuada o satisfactoria para la implementación en el reconocimiento de texto manuscrito continuo en el idioma aymará, al lenguaje del computador?

Tipo de Investigación: Descriptiva

Unidad de Observación: Reconocimiento automático de texto manuscrito continuo en el idioma aymara.

Variables: V_1: Modelos Estadísticos

V_2: Reconocimiento automático de texto manuscrito



Fuente: [Velásquez L.; Datos Propios]

Concepto: Reconocimiento automático de texto manuscrito continuo, donde se obtendrá como resultado un texto en lenguaje del computador, a partir de un texto escrito en aymara manuscritamente, considerando todos los atributos de este idioma aymara.

Indicador: Numérico

Instrumento: Probabilidades

Valor:

	Promedio en Reconocimiento (porcentaje)
Malo	0(%) - 50(%)
Regular	51(%) - 65(%)
Bueno	66(%) - 84(%)
Excelente	85(%) - 100(%)

1.5 OBJETIVO DE ESTUDIO

Investigar las bases teóricas sobre los modelos estadísticos para que nos lleven a una aplicación de reconocimiento de texto manuscrito continuo en lengua aymará.

La importancia de esta investigación puede ayudarnos a solucionar problemas de gran envergadura en las instituciones publicas y privadas ya que en la actualidad en nuestro país se tiene inmensos volúmenes de información escritas a mano que no están procesadas en el computador, como por ejemplo en identificación personal que otorga carnet de identidad que existe documentos que no se ha procesado digitalmente, libretas escolares, libretas de familia, cheques bancarios, etc.

1.6 JUSTIFICACION

La presente investigación sobre el tema planteado se puede abordar también en otros idiomas no solo en el idioma aymara ya que las características tienen mucha similitudes en otros idiomas excepto por ejemplo en el idioma chino con caracteres de escritura muy diferente, entonces es aplicable o se puede utilizar también en otros idiomas con los modelos estadísticos planteados. Con esta implementación los que serán beneficiados particularmente son las personas hablantes en el idioma aymará y quechua por las similitudes en escritura que tienen.

En el campo empresarial es de mucha utilidad una herramienta con estas características de reconocimiento de texto para su procesamiento al lenguaje computador ahorrando tiempo

y dinero a estas instituciones, se puede implantar también en el campo de reconocimiento de firmas que ayudaría mucho.

1.7 HIPOTESIS

H

El reconocimiento de texto manuscrito continuo a través de la utilización de los Modelos Ocultos de Markov (HMMs) permite un nivel de fiabilidad de ochenta y cinco por ciento, con relación a otras aplicaciones en el reconocimiento automático de texto manuscrito continuo.

1.8 OBJETIVOS

1.8.1 OBJETIVO GENERAL

Implementar la tecnología del reconocimiento de habla(RH), utilizando los modelos estadísticos; Modelos de Markov de capa oculta (HMMs), Autómatas Estocásticos de Estados Finitos (AEEF) y el modelado de lenguaje de N-gramas que permitan reconocer un texto manuscrito continuo en lengua aymará y su aplicabilidad en el lenguaje de maquina o computador.

1.8.2 OBJETIVOS ESPECIFICOS

- Plantear las bases teóricas para el desarrollo de reconocimiento de texto manuscrito continuo (RTM) mediante, HMMs, AEEF y el modelado de lenguaje de N-gramas
- Diseñar un modelo en reconocimiento de texto manuscrito en lengua aymará en base a las teorías estadísticas como, HMMs, AEEF y el modelo de lenguaje de N-gramas.
- Implementar el reconocimiento óptico de caracteres (OCR) para su procesamiento de la imagen off-line, conjuntamente con los modelos planteados y posterior reconocimiento de la escritura al lenguaje de maquina.

1.9 LIMITES Y ALCANCES

EL reconocimiento de texto manuscrito continuo aplicando los modelos estadísticos es el estudio principal a ser alcanzado. Específicamente los alcances a ser abordados son los siguientes:

- Modelos de Markov de capa oculta (HMMs) para el reconocimiento de caracteres manuscritos continuos aislados.
- Estudio minucioso de una característica en particular para luego ser generalizado para todo el alfabeto común.
- El reconocimiento óptico de caracteres (OCR), que nos permitirá procesar imágenes escaneadas de texto manuscrito en un formato procesable por el ordenador(ASCII), ingreso de datos de manera off-line por escáner .
- Autómatas Estocásticos de Estados Finitos (AEEF), para el proceso de reconocimiento del texto.

1.10 METODOLOGIA

1.10.1 METODOS Y MEDIOS DE INVESTIGACION CIENTIFICA

- **Análisis y Síntesis.** El vocablo análisis puede poseer distintos significados dependiendo de la disciplina en que se aborde, por ejemplo en términos de informática e ingeniería se tiene, análisis estructural, análisis de sistemas, análisis de conjunto, análisis fundamental, análisis FODA, análisis del entorno, y otros.

  El análisis en términos generales se refiere a la descomposición de un todo en sus distintos elementos, con el fin de estudiar estos de manera separada, luego en un proceso de síntesis se debe integrar todos esos elementos. En esta investigación el marco general es el estudio de las herramientas estadísticas, los elementos son modelos estadísticos señalados anteriormente, sintetizando para obtener el resultado en el objetivo planteado.
- **Inducción y Deducción**. La inducción se refiere a la generalización de una observación, razonamiento o conocimiento establecido a partir de casos particulares. [Geneviéve G.; 2003]

  La deducción es la aplicación de teorías genéricas a situaciones particulares.[Caraballoso K.; 2005]

  EL propósito en este tema del reconocimiento de texto manuscrito

Continuo, se plantea el reconocimiento automático del texto en un marco general, par luego llegar a lo particular que seria la herramienta a manejar para lograr este objetivo.

- **Abstracción y Concreción**. La abstracción es un proceso de suma importancia para la compresión del objeto, mediante ella se destaca la propiedad o relación de las cosas y fenómenos. No se limita a destacar y aislar alguna propiedad y relación del objeto asequible a los sentidos, si no que trata descubrir el nexo esencial oculto e inasequible al conocimiento empírico. [López J.; 1984].

  Lo concreto es la síntesis de muchos conceptos y por consiguiente de las partes. Las definiciones abstractas conducen a la reproducción de lo concreto por medio del pensamiento. Lo concreto en el pensamiento es el conocimiento más profundo y de mayor contenido esencial. [Ochoa A.; 2005].
- **Modelación**. La modelación es crear una representación explicita del entendimiento que una persona tiene de una situación, o simplemente de las ideas que se tiene acerca de una situación. Puede expresarse a través de matemáticas, símbolos o palabras, pero es esencialmente una descripción de entidades y las relaciones entre ellas. Pede ser descriptivo o ilustrativo, pero sobre todo debe ser útil.

  Existen tres formas de modelado:
  - Icónico: versión a escala del objeto real y con sus propiedades relevantes mas o menos representadas.
  - Analógico: modelo con apariencia física distinta al original, pero con comportamiento representativo.
  - Analítico: relaciones matemáticas o lógicas que representen leyes físicas que se cree gobiernan el comportamiento de la situación bajo investigación. [Sanloz H.; 1998]

1.10.2 METODOS Y MEDIOS DE INFORMATICA

Con respecto a los medios de informática se consideraran los siguientes modelos.

- Modelos de Markov capa oculta HMMs, para el reconocimiento de caracteres
- Autómatas Estocásticos de Estados Finitos AEEF.
- Modelado de lenguaje de N-gramas, para el modelado léxico

- Reconocimiento óptico de caracteres (OCR), para la migración de datos
- Maquinas de Estados Finitos (MEF), modelado en las frases

Toda esta teoría se implementara en el lenguaje Matlab 7.0, por que esta herramienta maneja términos matemáticos y probabilidades, donde se hace uso frecuente de los comandos en Matlab.

1.11 APORTES

1.11.1 APORTE TEORICO

Analizando con respecto a esta propuesta de investigación de reconocimiento de texto manuscrito continuo, se puede observar desarrollos realizados al respecto en otros campos como mencionados anteriormente, esta propuesta se basa sobre los modelos estadísticos mencionados para el reconocimiento de texto manuscrito continuo, que esta recobrando alto interés, ya que estos modelos podrían utilizarse en este contexto que promete resultados satisfactorios.

Modelos ocultos de Markov para el análisis de patrones espaciales. Los modelos ocultos de Markov (HMM) constituyen una herramienta de modelización altamente flexible, inicialmente utilizada en el campo del reconocimiento automático del habla, que ha encontrado en los últimos años numerosas aplicaciones en áreas científico-técnicas muy diversas, aunque su utilización en ecología es aún escasa. [Rodríguez F.;2006]

Los autómatas estocásticos de estados finitos (AEEF), que coadyuvaran en el reconocimiento del texto y los modelos de N-gramas que ayudarían en el modelado de lenguaje empleado en reconocimiento de frases.

1.11.2 APORTE PRÁCTICO

Se espera probar con éxito con los modelos planteados anteriormente un reconocimiento de texto manuscrito (escritura a mano continua) de manera automática en el idioma aymara ya que este promete una investigación por sus atributos propios de este idioma como por ejemplo: Se tiene escrituras manuscritas como T'uruña, jark'aña que tiene una manera diferente de articular estas palabras en aymara, con respecto a la presente investigación se tomara en cuenta el apostrofe que se muestra en estos ejemplos como **T'** y **k'** que se articula del modo glotalizado, puede ser bilabial, alveolar, velar y post-velar que se considerara en este tema de investigación de escritura manuscrita continua en el idioma aymara.

2 MARCO DE REFERENCIA

2.1 LA LENGUA AYMARA

Tanto las palabras quechua y aymara para designar a las respectivas “etnias” son producto de la confusión que complicó a los españoles cuando trataron de comprender a este mundo tan peculiar que estaban destruyendo. Los antecesores de los actuales aymaras nunca supieron que se llamaban así. Los incas los llamaban collas, hasta que en 1559 Polo de Ondegardo los denominó "aymaras" a partir de la información lingüística obtenida en el Collao de una pequeña colonia de mitimaes "quechuas", pero que habían incorporado el lenguaje local y que se denominaban aymaras y provenían de los alrededores de Cuzco. Así se llamó "en español" al idioma cuyo real nombre era jaqi aru (significando humanidad y lengua respectivamente) y después le aplicaron ese nombre a quienes hablaban ese idioma, quienes se llamaban a sí mismos jaqi. Algo similar ocurrió con el quechua, cuyo nombre real es runasimi y significa algo parecido.

Tratando de clasificar los lenguajes de América, Greenberg define una familia idiomática "andino-ecuatoriana". En una de sus sub-familias agrupa al quechua y al aymara, separándolos del uru-chipaya. La relación que existe entre ambos es difícil de delimitar. Los incas, exponentes del quechua a la llegada de los españoles, provenían de los pukinas del Tiwanaku y probablemente llegaron al Cuzco hablando esa lengua entre los de la clase gobernante. Pero la lengua corriente en el Tiwanaku (y podría ser que también en el Imperio Wari) era un jaqi primitivo (proto-jaqi) y probablemente éste fue el lenguaje del ciudadano común incaico durante algún tiempo. Sin embargo, los incas fueron fuertemente

influenciados por los Chinchay de la costa (Pachakamaq), quienes hablaban el idioma que hoy se denomina quechua y podría ser que esto difundió el quechua hasta el punto de desplazar al proto-jaqi y no tardaría mucho en perderse el idioma especial de la clase privilegiada. Es decir, en algún momento el estado incaico era trilingüe y la masa laboral, más o menos bilingüe. Allí se habrían generado las numerosas coincidencias superficiales entre el aymara (descendiente del proto-jaqi) y el quechua, aunque difieren en la profundidad de la estructura gramatical. [Infoarica.;2004]

Las lenguas andino-ecuatorianas carecían de escritura bajada en grafemas, por lo que su "traducción" al alfabeto latino es bastante caótica. Aunque hay estándares definidos para el aymara, en diversos textos de expertos en materias no lingüísticas aparecen diferentes ortografías, usando, por ejemplo, la "c" en vez de la "k", la "q", la "qh" o la "ch", las que representan convenciones referentes a la pronunciación original. Para facilitar la lectura escribo araj en vez de arax porque así figura en textos de antropólogos eruditos pero no lingüistas. También acepto castellanizar el plural con una s final, la cual no existe en el idioma quechua ni en el aymara (en el primero la pluralidad se consigue con el sufijo kuna, y en segundo con el sufijo naka, como en patanaka). El aymara tiene sólo tres vocales: "a", "i", "u". No existe, curiosamente, un fonema para la "e" ni la "o". Además he omitido los acentos para los vocablos indígenas. Nótese que en aymara se acentúa vocalmente la penúltima sílaba. [Infoarica.;2004]

La frase aymara más célebre es aruskipasipxañakasakipuniraskispawa, la cual, según un lingüista, contiene 14 sufijos para decir "yo sé que es deseable y obligación de todos, incluyéndolos a Uds., que nos comuniquemos" o en buen castellano, "conviene dialogar". [Infoarica.;2004]

2.1.1 ALFABETO AYMARA

Después de muchos años de propuestas, reuniones y deliberaciones, por el decreto supremo 20227-DS del 9 de mayo de 1984 del gobierno boliviano, denominado único. Se puede considerar a este alfabeto como una síntesis de los alfabetos de yapita y Marykonll (Tabla 2.1) que eran sistemas de amplia aceptación por los lingüistas y los pocos aymaristas usuarios.

El alfabeto oficial aymara consta de veinte y seis (26) consonantes y tres (3) vocales.

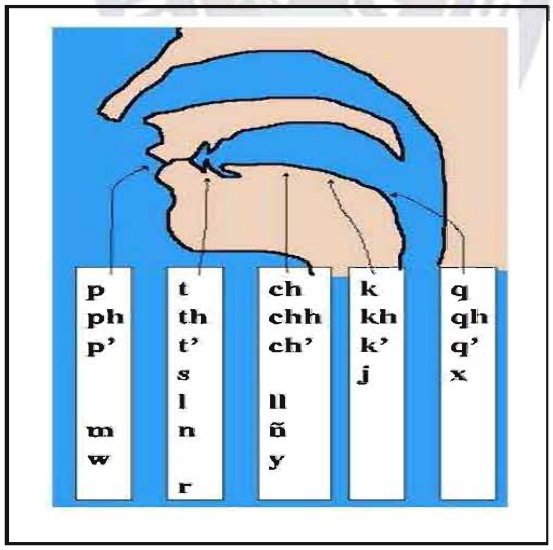
Es un sistema fonético, pues cada fonema considerado es representado por una única letra. Obviamente no es perfecto pues, en los diferentes dialectos, existen fonemas que no han sido representados en el mismo. Pero es conocido que en ninguna otra lengua del mundo existe una relación biunívoca entre fonemas y letras, por lo que convenciones ortográficas, como la de este alfabeto, son necesarias.

En la figura 2.1 se observa las 26 consonantes del alfabeto aymara clasificados en cinco(5) grupos, conforme el punto de articulación del aparato fonador humano. Estos 5 puntos de articulación usados para emitir lo fonemas considerados consonantes del alfabeto aymara son: bilabial, dental, palatal, velar y postvelar. [Pedraza J.; 2005]

En español las letras k y q son dos representaciones diferentes del fonema /k/. En aymara estas dos letras representan a dos fonemas diferentes. Esto puede ser notado, en la figura, por que los fonemas aymaras /k/ y /q/ son producidos en dos puntos de articulaciones diferentes: velar y postvelar, respectivamente.

Además en la tabla 2.1 es presentado en alfabeto aymara según su punto de articulación (conforme la figura 2.1) y según el modo de articulación (oclusivas simples, oclusivas, aspirada, oclusiva globalizada, etc.). [Pedraza J.; 2005]

Figura 2.1 Punto de articulación del aparato fonador humano



Fuente: [Pedraza jorge]

Tabla 2.1 Alfabeto Oficial sonorizado

CONSONANTES					
	Bilabial	**Dental**	**Palatal**	**Velar**	**Postvelar**
Oclusivas simples	p patas	t tita		k Lanka	q qaqa
Oclusivas aspiradas	ph phaxsi	th thaxa		kh khusa	qh qhach'u
Oclusivas glotalizadas	p' p'iqi	t' t'ula		k' k'utuña	q' q'urawa
Africada simple			ch chacha		
Africada aspirada			chh chhala		
Africada glotalizada			ch' ch'illiwa		
Fricativas		s saxra		j jach'a	x saxra
Laterales		l layqa	ll llij-lliju		
Nasales	m mamani	n nasa	ñ ñiq'i		
Vibrante		r saxra			
Pseudovocales	w wali		y yauri		
VOCALES					
Anterior	**Central**	**Posterior**	**Alargamiento vocálico**		
I	a	U	ï	ä	ü

Fuente: [Pedraza jorge]

2.1.2 Ortografía elemental

- Si en una palabra las vocales **i** y **u** son vecinas de los fonemas post-velares **x**, **q**,**qh** y **q'**, entonces son abiertos en */e/* y */o/* respectivamente. Sin embargo estas vocales abiertas son alófonos antes que vocales independientes. Elemplos: ñiq'i y q'urawa que son abiertos a *ñeq'e* e *q'orawa* respectivamente.

- No existen los diptongos. En lugar de ello se deben usar la semivocal **y** o la semiconsonante **w**. Ejemplo *wara wara* en lugar de *huara huara*.
- Cada palabra del aymara tiene su respectiva sílaba tónica. Casi todas las palabras del aymara son llanas. Ejemplos: ***na**ya*(yo), *ji**wa**sa*(tu y yo), *lura**ña**ni*(vamos a hacer). etc.
- Todas las propuestas alfabéticas del aymara, incluyendo la oficial, evitan el uso de tildes. Las acentuaciones gráficas en aymara son las dierisis que se usan en las vocales (silabas) largas. En este caso la tonicidad/alargamiento recae en estas sílabas. Ejemplos: escuche janiw ukham **sañäkiti** (No hay que decir así). Otros ejemplos: *sarä*(me voy), *janïpuniw* (jamás, nunca), etc.
- Hay una gran discusión en el uso de las letras **b**, **c**, **d**, **e**, **f**, **g**, **h**, **o**, **v**, z para los préstamos del español. Por ejemplo *bomba atómica* en aymara es fonetizado como *wumpa atumica*, *buenos dias* en *winus tiyas*, etc. Como debe escribirse? conforme es fonetizado usando el alfabeto aymara? o conforme aparecen escritos en español?.

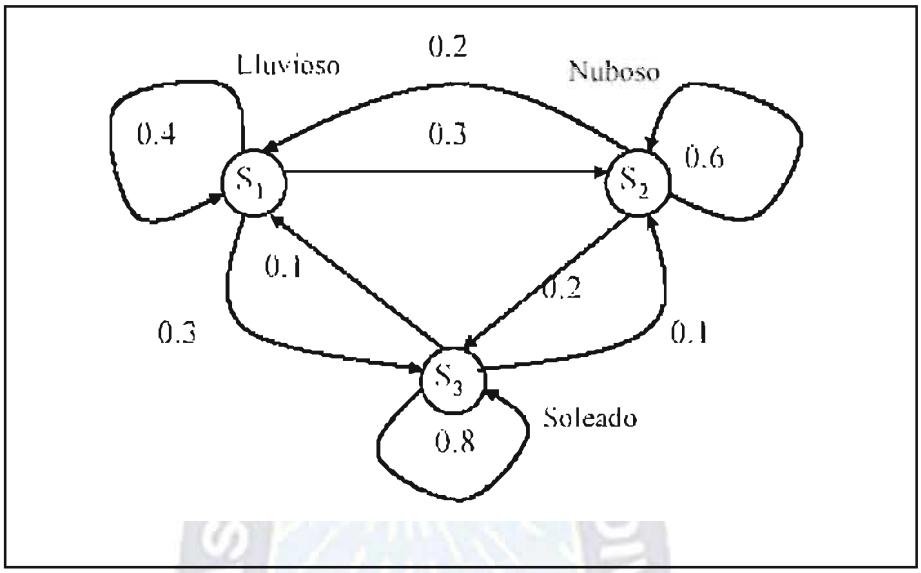
2.2 DEFINICION DE UN HMMs (HIDDEN MARKOV MODELS)

Los modelos de Markov describen un proceso de probabilidad el produce una secuencia de eventos o símbolos observables. Son llamados ocultos por que hay un proceso de probabilidad subyacente que no es observable, pero afecta la secuencia de eventos observados [Morgan ; 1991].

2.2.1 MODELOS OCULTOS DE MARKOV

Los modelos ocultos de Markov pueden ser vistos como el modelo de un proceso, el cual produce una secuencia de eventos acústicos perteneciendo a una unidad específica, o palabra, palabra en un vocabulario dado. Las variaciones entre las secuencias de observaciones de la misma clase, como la longitud de una palabra y pronunciación, son modeladas por la naturaleza del elemento estocástico de un HMMs.
Veamos un ejemplo en la figura 2.2 (estado del tiempo)

Figura 2.2 Ejemplo Autómata izquierda derecha



Fuente: [Bergasa L.; 2000]

Dado que hoy esta soleado.¿Cual es la probabilidad de la secuencia soleado, nuboso, lluvioso?

2.2.2 ALGORITMOS PARA MODELOS OCULTOS DE MARKOV (HMM)

- Algoritmo de Viterbi

  Calcula la trayectoria más probable en un HMM
- Algoritmo forward y backward

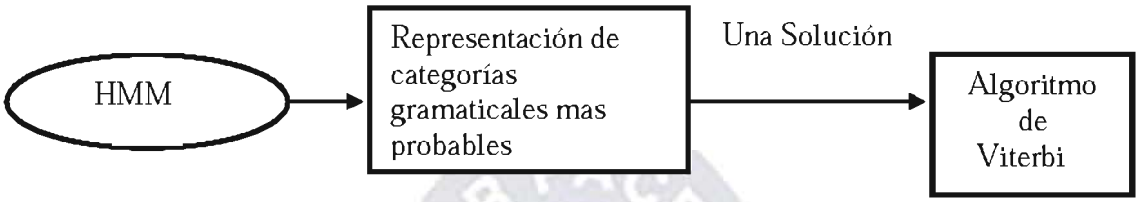
  Calculan la probabilidad de una secuencia de palabras.
- Algoritmo Forward-Backward (Baum-Welch)

  Estima las probabilidades asociados a un HMM.

2.2.2.1 ALGORITMO DE VITERBI

El algoritmo de Viterbi fue inicialmente desarrollado para encontrar, dada una secuencia de símbolos, la serie de transiciones más probable entre los estados de una cadena de Markov necesaria para producir dicha secuencia. Este problema es el equivalente markoviano al análisis sintáctico en una gramática regular estocástica. [Forney,2003]

El algoritmo de Viterbi es un caso particular del algoritmo de Programación Dinámica utilizado para encontrar un camino extremal en un grafo multietapa.



2.2.2.2 ALGORITMO DE VITERBI (forward)

$\sigma(n+1) = S_{1,n+1}$ mas probable que genero a $W_{1,n}$

$$\sigma(n+1) = \arg\max P(S_{1,n+1} / W_{1,n})$$

$$= \arg\max \frac{P(S_{1,n+1}, W_{1,n})}{P(W_{1,n})}$$

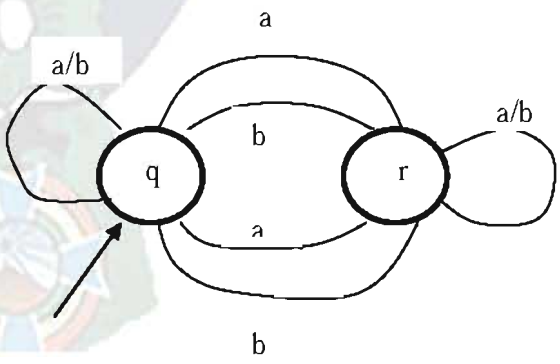
$$= \arg\max P(S_{1,n1}, W_{1,n})$$

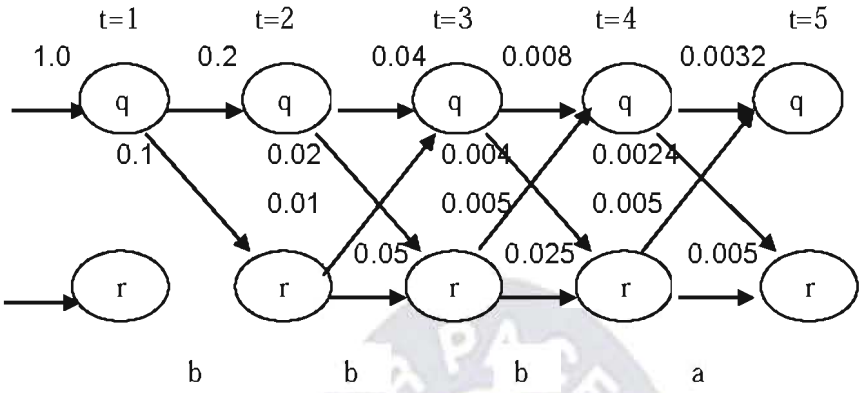
$\sigma_j(t) = S_{1,t}$ mas probable que genero a $W_{1,t-1}$ y termina en $S_t=S^l$

$$j = \arg\max P(\sigma_k(t)) P(S^k \rightarrow S^j)$$

Ejemplo:

S^l	t=1	t=2	t=3	t=4	t=5
$\sigma_q(t)$	q 1.0	qq 0.2	qqq 0.04	qqqq 0.008	Qrrrq 0.005
$\sigma_r(t)$	r 1.0	qr 0.1	qrr 0.05	qrrr 0.25	qrrrr 0.005
$\sigma_{1,t}$	q	qq	qrr	qrrr	qrrrq qrrrr





El algoritmo encuentra $S_{1,n+1}$ en un tiempo lineal.

[Rodríguez, F.; Bautista S.; 2006]

2.2.2.3 ALGORITMO DE VITERBI (backward)

$\gamma(n+1) = S_{1,n+1}$ mas probable que genero a $W_{1,n}$

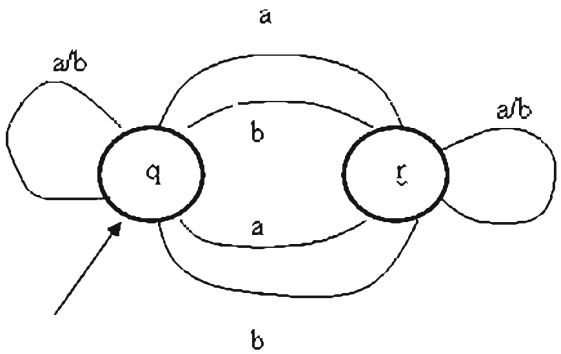
$\gamma(n+1) = \arg\max P(S_{1,n+1}, W_{1,n})$

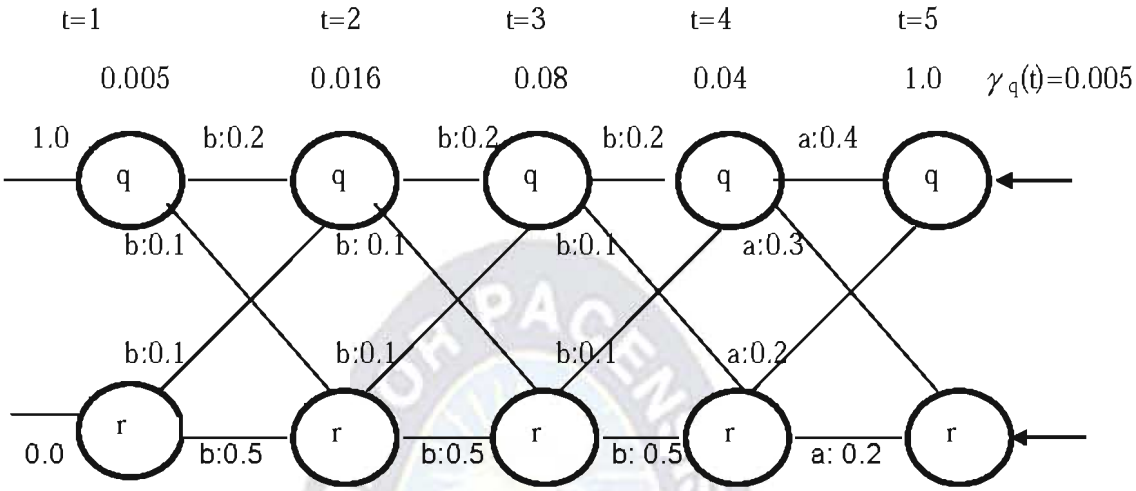
$\gamma_i(t) = S_{1,t}$ más probable que genero a $W_{1,t-1}$ y que comienza en $S_1=S^i$

$\gamma_i(t) = \arg\max P(S_1=S^i, S_{2,t}, W_{1,t-1})$

$j = \arg\max P(S^i \rightarrow S^k)\, P(\gamma_k(t))$

Ejemplo





[Rodríguez, F.; Bautista S.; 2006]

2.2.2.4 ALGORITMO FORWARD

$$P(W_{1,n}) = \sum_{i=1}^{ns} P(W_{1,n}, S_{n+1} = S^i)$$

$$\alpha_i(t) = P(W_{1,t-1}, S_t = S^i) \qquad t>1$$

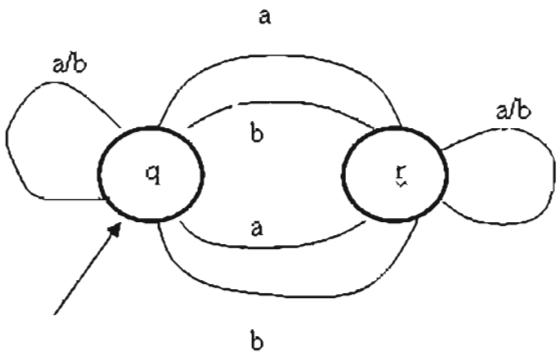
$$\alpha_i(t) = \begin{cases} 1 & i=1 \text{ comenzar en el estado } S^1 \\ 0 & \text{e.o.} \end{cases}$$

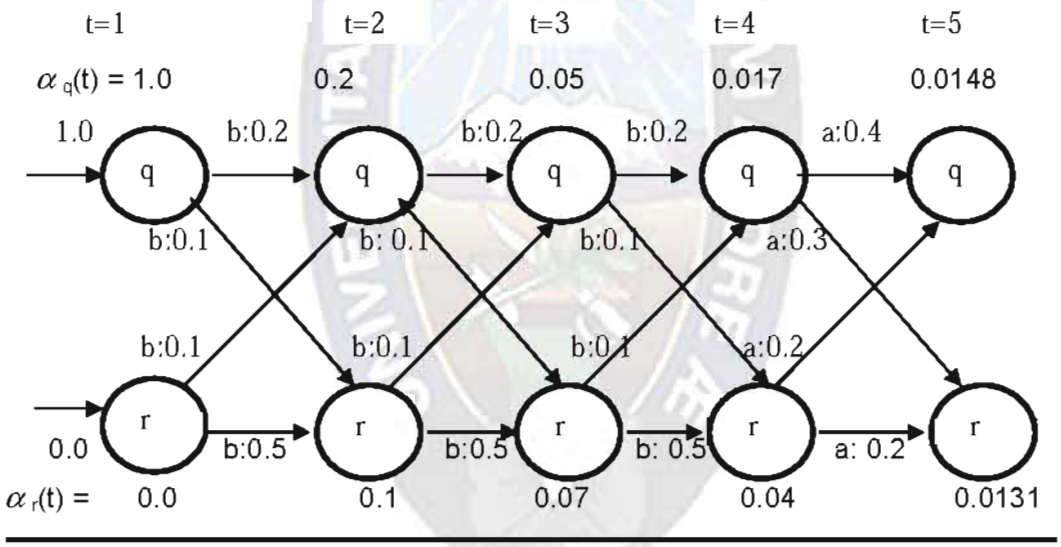
$$P(W_{1,n}) = \sum_{i=1}^{ns} \alpha_i(n+1) \quad ;$$

$$\alpha_i(t+1) = P(W_{1,t}, S_{t+1} = S^i)$$

$$= \sum_{i=1}^{ns} P(W_{1,t}, S_t = S^i, S_{t+1} = S^i)$$

$$= \sum_{i=1}^{ns} P(W_{1,t}, S_t = S^i, S_{t+1} = S^j)\, P(W_t, S_{t+1} = S^j / S_t = S^i) = \sum_{i=1}^{ns} \alpha_i(t) P(S^i \rightarrow S^j)$$



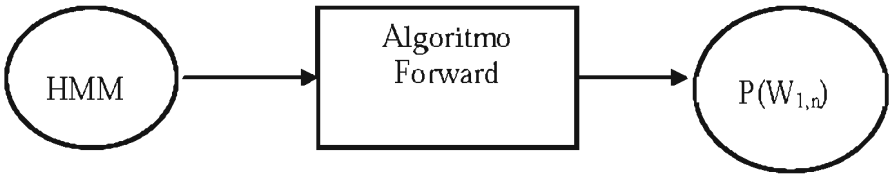


$P(W_{1,t})$= 1.0 0.3 0.12 0.057 0.0279

$$P(bbba) = \sum_{i=1}^{ns} \alpha_i(5)$$

[Rodríguez, F.; Bautista S.; 2006]

2.2.2.5 ALGORITMO BACKWARD



$$P(W_{1,n}) = P(W_{1,n} / S_1 = S^1)$$

$$\beta_i(t) = P(W_{1,n} / S_t = S^1) \rightarrow P(W_{1,n}) = \beta_1(1)$$

$$\beta_i(n+1) = P(\varepsilon / S_{n+1} = S^i) = 1$$

$$\beta_i(t-1) = P(W_{t-1,n,} / S_{t-1} = S^i)$$

$$= \sum_{j=1}^{ns} P(W_{t-1,n,}, S_t = S^j / S_{t-1} = S^i)$$

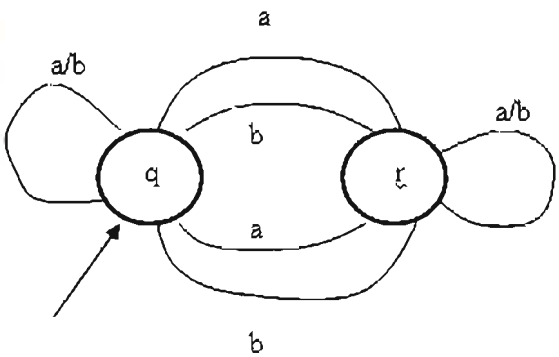
$$= \sum_{j=1}^{ns} P(W_{t-1,}, S_t = S^j / S_{t-1} = S^i)\, P(W_{t,n} / S_t = S^j)$$

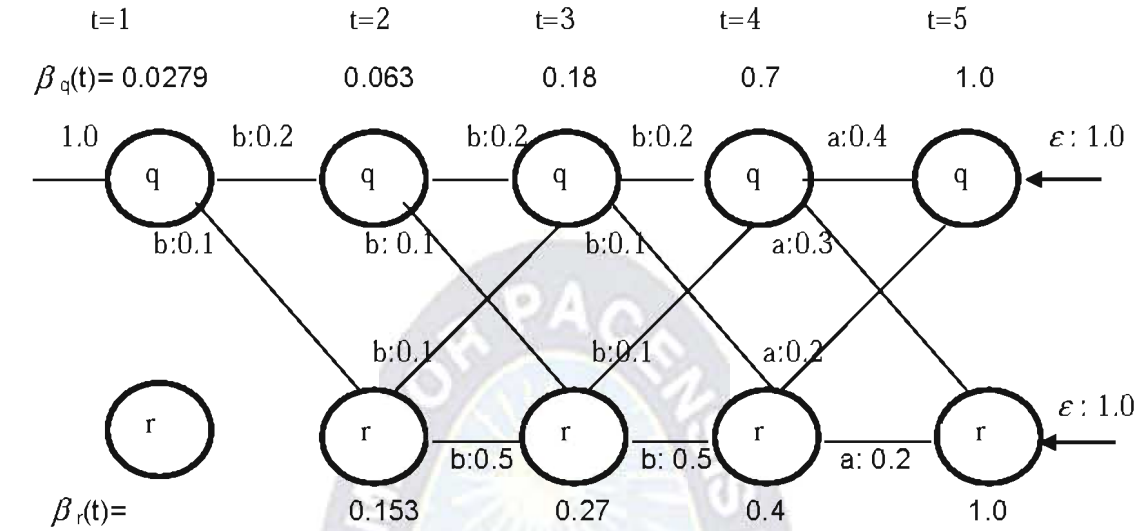
$$= \sum_{j=1}^{ns} P(S^i \rightarrow S^j)\, \beta_j(t)$$

Ejemplo

$$\beta_i(t=5) = P(\varepsilon / S_t = S^i) = 1$$

$$\beta_i(t-1) = \sum_{i=1}^{ns} P(S^i \rightarrow S^j)\, \beta_j(t)$$





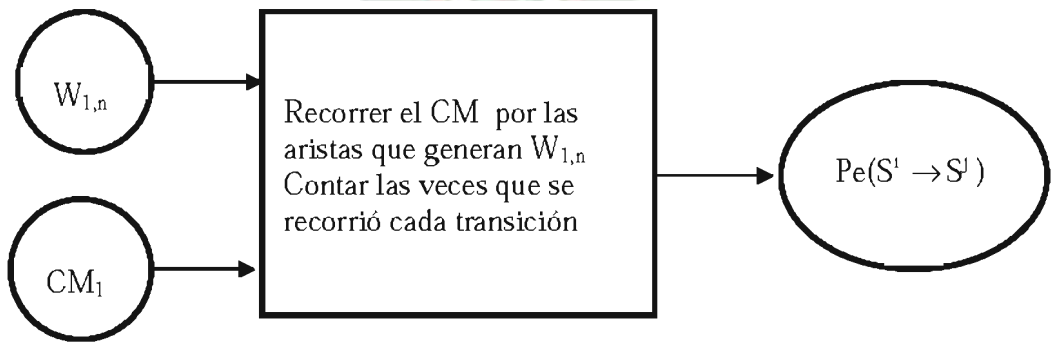
P(bbba) = $\beta_1(1)$

= 0.0279

[Rodríguez, F.; Bautista S.; 2006]

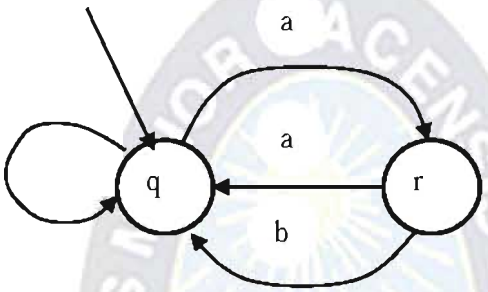
2.2.3 ENTRENAMIENTO DE LOS HMMs (HIDDEN MARKOV MODELS)

Esta tarea requiere además de métodos para llevar a cabo la estimación de estos parámetros a partir de estas muestras de entrenamiento. Los métodos más comunes para la estimación de dichos parámetros, consisten en establecer valores de los mismos de forma tal que maximicen la probabilidad de los modelos de generar tales secuencias de muestras de entrenamiento. Ésto es lo que se conoce como método de entrenamiento por *"máxima verosimilitud"*. Puesto que una solución analítica es inviable para estimar tales parámetros, éstos deberán serlo por algún método de *"descenso por gradiente"*. Un método muy común de reestimación iterativa es el algoritmo *"backward-forward"* o *"reestimación por Baum-Welch"*, el cual es una instancia del algoritmo EM . [Rodríguez, F.; Bautista S.; 2006]



$$Pe(S^i \rightarrow S^j) = \frac{C(S^i \rightarrow S^j)}{\sum_{h=1,m=1}^{ns,nw} C(S^i \rightarrow S^h)}$$

Ejemplo



Secuencia de entrenamiento : abbaababbaaa

	a	b	b	a	a	b	a	b	b	a	a	a

Secuencia de transiciones: $q \rightarrow r \rightarrow q \rightarrow q \rightarrow r \rightarrow q \rightarrow q \rightarrow r \rightarrow q \rightarrow q \rightarrow r \rightarrow q \rightarrow r$

$C(q \xrightarrow{a} r) = 5$ $\quad Pe(q \xrightarrow{a} r) = 5/8$

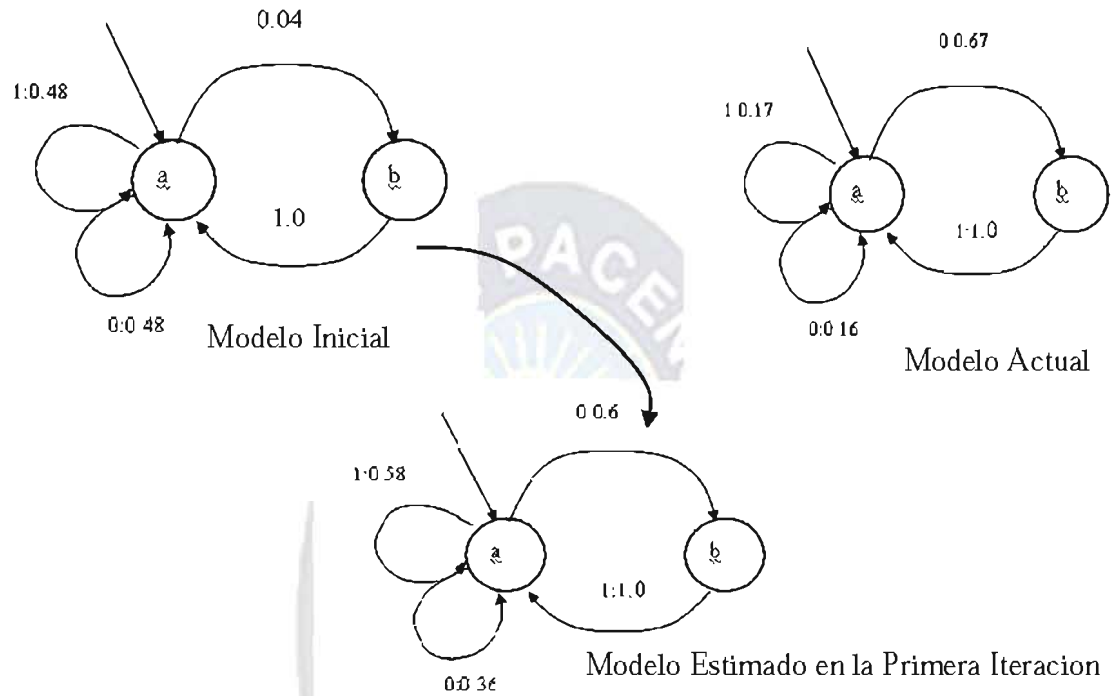
$C(q \xrightarrow{b} q) = 3$ $\quad Pe(q \xrightarrow{b} q) = 3/8$

$C(r \xrightarrow{a} q) = 2$ $\quad Pe(r \xrightarrow{a} q) = 2/4$

$C(r \xrightarrow{b} q) = 2$ $\quad Pe(r \xrightarrow{b} q) = 2/4$

ENTRENAMIENTO DE UN HMM

Ejemplo

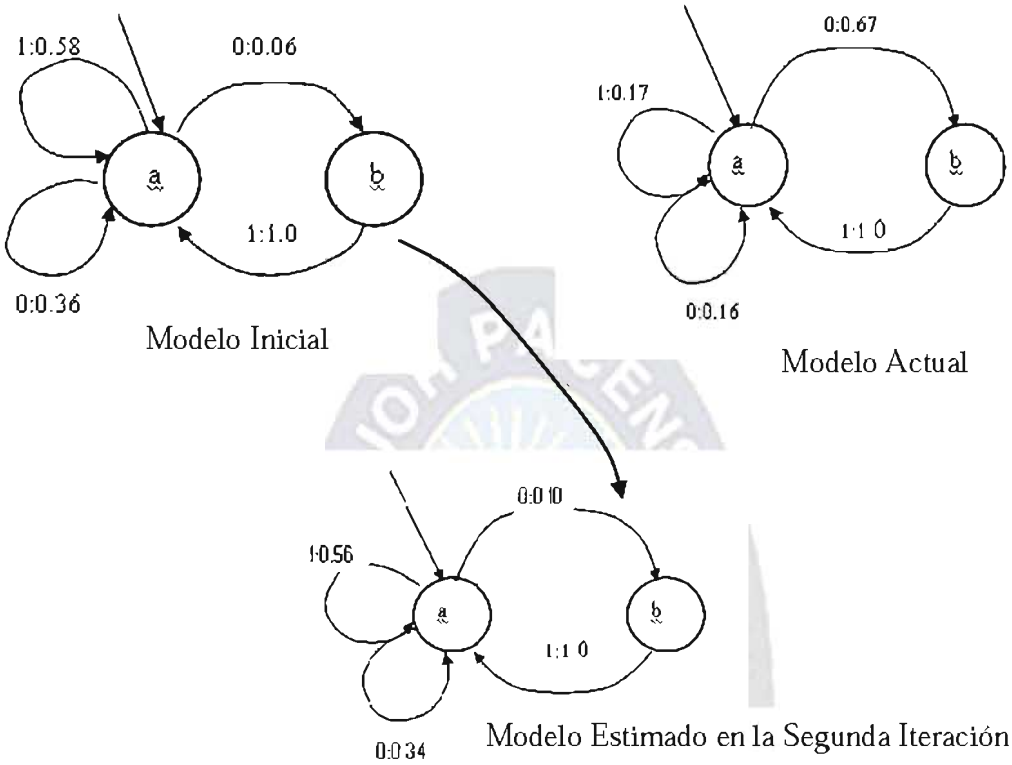


$W_{1,n} = 0.1011$

$S_{1,n}$	P(S,W)	a-0-b	b-1-a	a-0-a	a-1-a
ababaa	0.00077	0.00154	0.00154	0	0.00077
abaaaa	0.00442	0.00442	0.00442	0.00442	0.00884
aaabaa	0.00442	0.00442	0.00442	0.00442	0.00884
aaaaaa	0.02548	0	0	0.0597	0.09489
Total	0.03509	0.01038	0.01038	0.05970	0.09489
$P(S_i S_j)$	0.035	0.06	1	0.36	0.58

$$\sum a \xrightarrow{w} s = 0.01038 + 0.05970 + 0.09489 = 0.164997$$

$$\sum b \xrightarrow{w} s = 0.01038$$



$W_{1,n}$ = 01011

$S_{1,n}$	P(S,W)	a-0-b	b-1-a	a-0-a	a-1-a
ababaa	0.00209	0.00418	0.00418	0	0.00209
abaaaa	0.00727	0.00727	0.00727	0.00727	0.01454
aaabaa	0.00727	0.00727	0.00727	0.00727	0.01454
aaaaaa	0.02529	0	0	0.05058	0.07587
Total	0.04192	0.01872	0.01872	0.06512	0.10704
$P(S_i, S_j)$		0.10	1	0.34	0.56

$$\sum a \xrightarrow{w} s = 0.01872 + 0.06512 + 0.10704 = 0.19088$$

$$\sum b \xrightarrow{w} s = 0.01872$$

[Rodríguez, F.; Bautista S.; 2006]

2.2.4 CADENAS DE MARKOV

Una cadena de Markov $q = \{q_t\} t \in N$ es un proceso estocástico de markov discreto. Un proceso estocástico se llama de markov si conocido el presente, el futuro no depende del pasado, esto quiere decir, que dada una variable estocástica q_{t-1} que denota el estado del proceso en el tiempo t-1, entonces la probabilidad de transición en el momento t se define como:

$$P[q_t = \sigma_t / q_{t-1} = \sigma_{t-1}]$$

Formalmente, una cadena de markov se define como (Q,A), donde Q={1,2,……,N} son los posibles estados de la cadena y $A=(a_{ij})_{r \times m}$ es una matriz de transición de estados en el modelo.

Si $A(t)=a_{ij}(t)_{r \times m}$ es independiente del tiempo entonces el proceso se llama homogéneo y la probabilidades de transición de estados son los de la forma $a_{ij}(t)=P[q_t=j/q_{t-1}=i]$ son las siguientes probabilidades.

$$0 \leq a_{ij} \leq 1 \qquad 1 \leq i, j \leq N$$

$$\sum_{j=1}^{N} a_{ij}=1 \qquad 1 \leq I \leq N$$

La condición fundamental de que sea una cadena de markov establece que las probabilidades de transición y emisión depende solamente del estado actual y no del pasado, esto es $P[q_t = j / q_{t-1} =, q_{t-2} = k, \ldots\ldots\ldots] = P[q_t = j / q_{t-1} = i] = a_{ij}(t)$

Por ejemplo, la secuencia resultante de lanzar una moneda k veces es un proceso estocásticos. Los estados del sistemas son dos: cara y sello, estos a su vez conforman el espacio muestral, son el conjunto de todos los estados posibles del ejemplo.

2.2.5 TIPOS DE MODELOS DE MARKOV.

Los tipos de modelos de markov se observa en la tabla 2.2 .

Tabla 2.2 Tipos de Modelos de Markov

	Estado observable	**Estado oculto** (observaciones)
No hay acciones (sólo transiciones estocásticas)	Cadenas de Markov (Markov Chains)	Modelos Ocultos de Markov (HMMs)
Hay acciones (que producen transiciones estocásticas)	Procesos de Decisión de Markov (MDPs)	Procesos de Decisión de Markov Parcialmente Observables (POMDPs)

- **cadenas de markov**. En una cadena de Markov el estado es completamente observable (conocido en todo momento), y no existen acciones (no hay toma de decisiones, equivale a una sola acción)
- **modelos ocultos de markov** (HMMs). En un modelo oculto de Markov el estado es parcialmente observable, no se conoce pero existen observaciones y no existen acciones, no hay toma de decisiones y equivale a una sola acción.
- **procesos de decisión de markov**(MDPs). En un proceso de decisión de Markov el estado es completamente observable (conocido en todo momento), y existen diferentes acciones (un modelo de transición para cada acción p(s'/Saa)).
- **procesos de decisión de markov parcialmente observable**(POMDPs). En un proceso de decisión de Markov parcialmente observable el estado es parcialmente observable, no se conoce pero se dispone de observaciones y existen diferentes acciones, un modelo de transición para cada acción p(s'/s,a) [Bergasa L.; 1998]

2.2.6 DEFINICION DE UN HMM CONTINUO

Generalmente un HMM continuo M es una maquina de estados finitos que puede ser definida de la siguiente manera, por la séxtupla (Q,I,F,X,a,b) donde:

- Q es un conjunto finito de estados, que incluye el estado inicial $I \in Q$ y un estado final $F \in Q$.

- X es un espacio real: $X \subseteq \Re^d$
- a: (Q-{F}) x (Q-{I}) $\Rightarrow$ [0,1] es una función de distribución de probabilidad de transición entre estado.

$$\sum a(q_i, q_j) = 1 \qquad \forall q_i \in (Q-\{F\})$$

- b: (Q-{I,F}) x X $\Rightarrow$ [0,1] es una función de densidad de probabilidad de emitir un vector $x \in X$ en un estado $q_i \in Q$

2.3 AUTOMATAS ESTOCASTICOS DE ESTADOS FINITOS (AEEF)

Se define como una maquina de estados finitos definida por la séxtupla $(Q, \sum, R, q_0, P, F)$, donde:

- Q es un conjunto finito de estados
- $\sum$ es un conjunto finito de símbolos de entrada.
- $R \subset Q \times \sum \times Q$ es un conjunto de transiciones de la forma (q,d,q') para d $\in \sum$ y $q,q' \in Q$
- q_0 es el estado inicial
- P función de probabilidad de transición
- F función de probabilidad de estado final

2.3.1 AUTOMAS FINITOS

Los autómatas finitos son modelos útiles en software y hardware algunos ejemplos:

- Software par diseñar y verificar el comportamiento de circuitos digitales.
- Analizador léxico de un compilador (identifica y clasifica las palabras del lenguaje: identificadores, literales, operadores)
- Software para explorar grandes corpus de texto(colección de paginas web), en busca de palabras, frases, etc.
- Software para verificación de sistemas con un número finito de estados diferentes: protocolos de comunicación o de seguridad. [Mico M.; Forcada M.; Calera J.; 2006]

2.3.2 AUTOMATAS ESTOCASTICOS

Del mismo modo en que se extienden las gramáticas para obtener gramáticas estocásticas, se pueden extender los reconocedores para obtener reconocedores estocásticos. En concreto, para las gramáticas estocásticas de topo 3 o regulares, los reconocedores designados serán los autómatas finitos estocásticos, que se derivan de los autómatas finitos añadiendo probabilidades a las transiciones.

Un autómata finito estocástico (AFE) es una quíntupla (Q,V, δ ,I,F), donde Q es un conjunto finito de estados y V el conjunto de s terminales. ∂ : QxExQ $\rightarrow$[0,1] es una función parcial que asigna las probabilidades a todas las posibles transiciones. I:Q $\rightarrow$[0,1], F:Q $\rightarrow$[0,1] definen para cada estado su probabilidad de ser inicial y su pertenencia o no al conjunto de estados finales, respectivamente. Para asegurar la estocasticidad, la suma de probabilidades de las transiciones que salen de un mismo estado debe ser igual a la unidad:

$$\forall q \in Q, \quad \sum_{q_j \in Q,\, a \in V} \delta(q,a,q_i) = 1 \quad \wedge \quad \sum_{q \in Q} I(q) = 1$$

Se pueden dar toda una serie de definiciones similares a las formulas para gramáticas estocásticas:

Si el AFE es no-restringido (la probabilidad de una transición es independiente de las demas transiciones) la probabilidad de aceptar una cadena $\alpha = \alpha_1 \alpha_2 \ldots\ldots\ldots \alpha_n$ siguiendo la secuencia de estados $q_0, q_1, \ldots\ldots, q_n$ se escribe como:

$I(q_0)\ \delta(q_0, \alpha_1, q_1) \ldots\ldots\ldots\ldots\ \delta(q_{n-1}, \alpha_n, \alpha n) F(q_n)$

Dado que para una determinada cadena α pueden haber varios caminos a través del autómata que lleguen al estado final, cada uno con su probabilidad $P_c(\alpha)$, la probabilidad de generación de una cadena se puede definir también mediante la aproximación,

$$P(\alpha) = \sum_{\forall c} P_c(\alpha) \quad \text{(activa)} \quad \text{o bien} \quad p(\alpha) = \max_{\forall c} \{P_c(\alpha)\} \quad \text{(maximal)}$$

Al igual que las gramáticas el lenguaje aceptado por un AFE esta formado por las cadenas que a partir del estado inicial llevan a un estado final, con su probabilidad de aceptación. Se

dice que los AFE son equivalentes si aceptan el mismo lenguaje ponderado (mismas cadenas y mismas probabilidades).

Dada una gramática estocástica regular, existe un AFE que acepta un lenguaje idéntico al generado por la gramática (pero lo contrario no es cierto). [Gonzáles ; 1978]

2.4 MODELO DE LENGUAJE

Un modelo de lenguaje es un mecanismo para definir la estructura del lenguaje, es decir, para restringir adecuadamente las secuencias de unidades lingüísticas más probables.
En general son útiles en aplicaciones que exhiban una sintaxis y/o semántica completa, un buen modelo de lenguaje solamente debería aceptar frases correctas y rechazar aquellas secuencias de palabras incorrectas.

2.5 RECONOCIMIENTO OPTICO DE CARACTERES

En el reconocimiento óptico de caracteres hace énfasis en la implementación de un sistema para el reconocimiento de caracteres manuscritos aislados, basados en la tecnología del reconocimiento del habla. Ya que esta tecnología también se basa en los modelos ocultos de markov. Estos modelos son adecuados para la segmentación en el proceso de reconocimiento de señales.
Se hace un reconocimiento de caracteres aislados individuales, por ejemplo el tema propuesto como investigación del reconocimiento de texto manuscrito continuo.
En el OCR ("Ópticas Carácter Recognition"), que coadyuvara en un reconocimiento final

El OCR e ICR son herramientas muy usadas en informática. Estos son métodos automatizados de la recogida de datos ampliamente utilizados para el proceso de alta velocidad de gran capacidad de formas idénticas.
Este es un software especializado para dar vuelta a imágenes en el texto editable, nos proporcionan la capacidad de dar vuelta a documentos de la copia dura en formatos reutilizables de su opción tales como archivos del MS palabra, los archivos pdf, las paginas de hml, txt , listos para el almacenaje y la recuperación rápidos organizados, y simple. El OCR se puede también utilizar para dar vuelta a archivos imagen-basados del pdf nuevamente dentro del texto editable. El OCR es un sistemas de la exploración y de la proyección de imagen, la capacidad de dar vuelta a imágenes de caracteres maquinas-impresos en caracteres legibles por la maquina. El software del OCR puede conocerse una

variedad amplia de fuentes en diversos idiomas, utilizamos el reconocimiento del texto del OCR para convertir datos explorados de la imagen en los archivos de textos digitales para la gerencia del documento, el publicar de procesamiento de textos los usos e indexación de direcciones automáticas (exploración borrosa), de escritorio. Las facultades más grandes del OCR con su capacidad de incorporar datos en una base de datos rápidamente y su exactitud de lectura demostrada. [Green P.; 2000]

El OCR puede leer virtualmente cada clase de texto impreso por ejemplo:

- Dígitos
- Símbolos
- Letras mayúsculas
- Letras minúsculas
- Letras acentuadas
- Puntuación

Se puede proporcionar métodos de proceso del documento del OCR en varios formatos por ejemplo:

- Cheques con el OCR
- Con texto completo con el OCR
- El poner en un índice usado el OCR
- Texto manuscrito con ICR

2.5.1 RECONOCIMIENTO DE CARACTERES INTELIGENTE (ICR)

Esta tecnología avanzada de la exploración puede traducir una variedad amplia de las fuentes y del tipo estilos impresos de las fuentes de papel al texto electrónico, permite a sistemas de la exploración y de la proyección de imagen dar vuelta a imágenes de caracteres impresos a mano en caracteres legibles por la maquina.

Los sistemas ICR llevan a los OCR un paso adelante utilizando programas de cómputo para aplicar pruebas de inteligencia lógica a los caracteres escaneados para convertirlos de manera más confiable en información más legible para la computadora.

Los sistemas de ICR aplican reglas de ortografía, gramática y contexto para escanear los textos a fin de efectuar evaluaciones "inteligentes" sobre la interpretación correcta de la información. Esto permite una conversión mucho más precisa de los textos escaneados de la que realizan los sistemas OCR más simples, especialmente con el texto manuscrito.

Los programas ICR requieren computadoras rápidas y poderosas para desempeñarse de manera eficiente: Los sistemas ICR de alta confiabilidad solo estuvieron disponibles a mitad de la década de 1990 con el desarrollo de productos computacionales, económicos y poderosos.

A medida que se vuelvan más confiables los sistemas ICR, se incrementarán sus aplicaciones electorales. Son particularmente apropiados para capturar información de formatos. También se está evaluando su capacidad para capturar números manuscritos de las papeletas que utilizan sistemas electorales más complejos, como el de voto alternativo o el de voto único transferible. A la fecha, los sistemas automatizados de captura de información no han sido utilizados para estos sistemas electorales debido a la complejidad de la tarea. [Green P.; 2000]

2.6 PREPROCESAMIENTO

Este módulo tiene por objetivo minimizar las variaciones que involucran los diferentes estilos de escritura (y dentro del mismo estilo también) dependientes, entre otros factores, del escritor. Entre estas variaciones se encuentran por ejemplo los diferentes grados de inclinación del trazo (*"slant"*), las relaciones de altura entre clases de caracteres como ascendentes (ej. "b") o descendentes (ej. "g") y normales (ej. "a"), el grosor del trazo, el ancho de las letras, etc. Estas peculiaridades son irrelevantes a los fines del reconocimiento y tienden a oscurecer la identidad de los caracteres, aunque en otras situaciones, por ejemplo *"verificación de firmas"*, pueden tener más importancia. Los atributos de estilo a normalizar que consideraremos y que solo interesan a caracteres manuscritos aislados se enumeran a continuación: [Doménech J.; 2000]

- **Nivel de ruido:** Que puede aparecer en el proceso de digitalización o que sea intrínseco de la imagen.
- **Inclinación o "Slant":** Es el ángulo del trazo respecto a la vertical.
- **Tamaño:** La altura de la letra varía entre diferentes escritores para una tarea dada, y para un escritor dado entre diferentes tareas.

Cada uno de estos atributos de estilo se caracteriza por medio de una serie de parámetros que varían entre diferentes instancias de un carácter. El proceso busca en definitiva, eliminar los efectos de tales parámetros sobre la escritura.

2.6.1. FILTRADO DE RUIDO

Normalmente en la fase de digitalización de imágenes se introducen pequeñas manchas o componentes aisladas de píxeles negros. El simple hecho de quitarlas porque se consideran "ruido", se justifica por el hecho de que la permanencia de las mismas podría tener repercusiones importantes en etapas posteriores.

Tales manchas o componentes aisladas tienen efectivamente una gran repercusión sobre el cálculo de la caja mínima de inclusión, ocasionando que las operaciones que dependen de esta caja, se vean seriamente afectadas por las manchas. La figura 2.3 muestra un ejemplo de preproceso sin y con filtrado de manchas.

Como se puede apreciar los casos mínimas de inclusión de ambas situaciones encierran áreas de imagen muy diferentes.

Justificada ya la razón del filtrado de ruido, la primera función del módulo de preproceso consiste en eliminar dichas manchas de la imagen. Para ello se busca en toda la imagen conjuntos aislados de píxeles conectados, que son luego eliminados si su talla no supera un umbral predeterminado.

Una vez suprimido el ruido, se hace el análisis siguiente extracción de la imagen propia del carácter (píxeles negros) del resto de su fondo (píxeles blancos) ". [Doménech J.; 2000]

Figura 2.3 Ejemplo de preproceso sin (derecha) y con (izquierda) filtrado de manchas para la letra manuscrita "d".

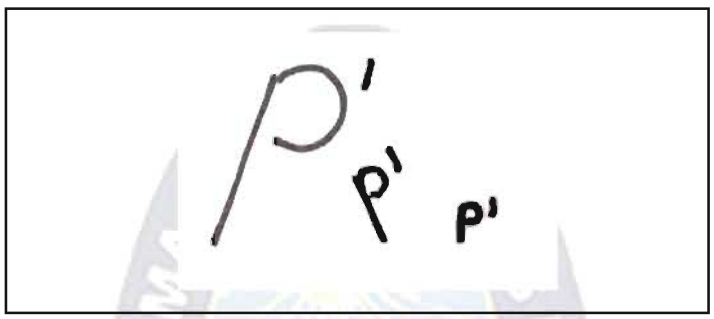
Fuente: [Elaboración Propia]

2.6.2. CORRECCION DE "Slant"

Una de las características del texto manuscrito es que no siempre se presenta en posición vertical, como el texto impreso, sino que cada escritor suele darle un grado de inclinación característico de su estilo de escritura. La figura 2.4 muestra algunos ejemplos de letras

manuscritas aisladas, en la figura se muestra el carácter "p" con una apostrofe acompañada, este carácter propio del aymara se planteara en estudio mediante los modelados propuestos.

Figura 2.4 Ejemplos de 3 estilos de escritura de la letra minúscula " p' " realizados por diferentes escritores, caracterizados por distintos grados de inclinación.



Fuente: [Elaboración Propia]

La segunda fase de preproceso está encargada de realizar una corrección del *"slant"* sobre la imagen. El método aplicado consiste básicamente en la convolución de la imagen *fentr*(*x*, *y*) con los operadores de *"Sobel"* horizontal y vertical:

$$GS_h = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \qquad GS_V = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$$

para obtener las imágenes resultantes correspondientes *h*(*x*, *y*) y *v*(*x*, *y*). A continuación se calcula el ángulo de fase del gradiente en cada punto de la imagen (tomando como referencia el eje vertical) como:

$$\text{fase}(x, y) = 90 - \arctan\left(\frac{v(x, y)}{h(x, y)}\right)$$

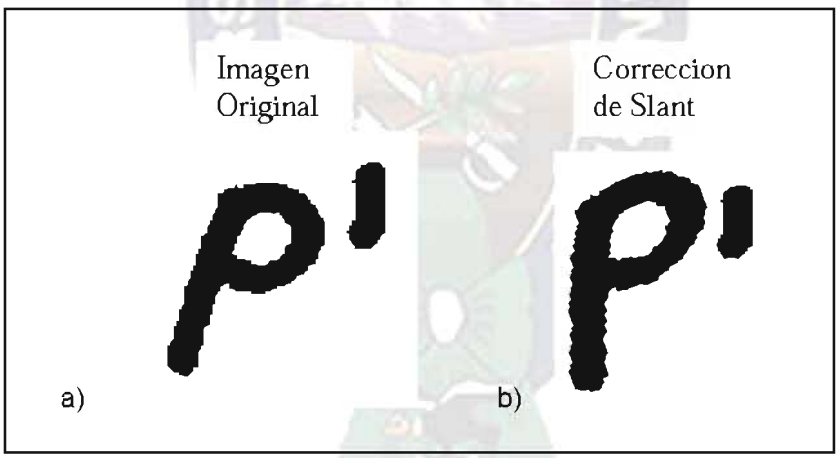
Se genera un histograma de frecuencias de ángulos de fase HIST(fase(*x*, *y*)) mediante la contabilización del número de apariciones de cada magnitud de ángulo en fase(*x*, *y*) .

Con objeto de priorizar los ángulos que están más cercanos a la vertical respecto aquellos que están casi horizontales, se hace necesario aplicar un suavizado a dicho histograma. Para ello se probaron varias funciones de suavizado entre las que el triángulo unidad:

$$F_s(\alpha)=\left(1-\frac{|\alpha|}{90}\right) \qquad \alpha \in [-90,90]$$

demostró ser una de las más convenientes a la vista de los resultados finales.
Finalmente, del histograma suavizado HIST$s(\alpha)$ se computa la media, la que es tomada como valor del parámetro de *"slant"* dominante de la imagen. Aplicando una transformación de desplazamiento acorde a este parámetro, se corrige la inclinación vertical de la imagen ver figura 2.5. [Doménech J.; 2000]

Figura 2.5: Resultados del preproceso para corrección de *"Slant"* de una imagen con la letra manuscrita "p' ", a) imagen original con inclinación a la derecha, b) imagen corregida con respecto a la vertical.



Fuente: [Elaboración Propia]

2.7 EXTRACCION DE CARACTERISTICAS

El objetivo de la fase de extracción de características es la obtención de secuencias de vectores de características que describan los objetos (caracteres manuscritos aislados en este caso) de forma tal que se diferencien entre sí y puedan ser distinguidos evitando tanto como sea posible la información redundante.

Es deseable experimentar con un conjunto de características que sean fácilmente obtenibles a un costo en recursos relativamente bajo, que sean independientes entre sí y cuyo esquema de extracción sea adecuado para su utilización conjunta con HMM. Se consideran un conjunto de características los requerimientos sustentados sobre mediciones geométricas: niveles de gris, derivadas horizontales, derivadas verticales, pendiente local y coeficiente de correlación.

2.8 ESQUEMA PROBABILISTICO DE CARACTERES: HMMs(Hidden Markov models)

Dada una secuencia de vectores (reales) de características $x = (x_1 x_2 \quad x_N)$ con $x_i \in \Re^d$ la probabilidad que corresponda a un determinado carácter "*c*", viene expresado por la regla de Bayes:

$$P(c/X) = \frac{p(X/c)\,P(c)}{p(x)}$$

donde:

- $P(c)$: es la probabilidad a priori de la clase de carácter *c*.
- $p(x)$: es la densidad de probabilidad de la secuencia de vectores (reales) de características x.
- $p(x|c)$: es la densidad de probabilidad condicional de una secuencia de vectores de característica x para una clase de carácter *c*.

Para estimar la probabilidad $p(x|c)$ se utilizan los HMMs continuos. Básicamente los HMM son máquinas de estados finitos que, en este caso, se utilizan para modelar la secuencias de vectores de característica derivados de instancias de estos caracteres. Para una definición formal de HMM.

Se asume que cada estado de un HMM genera vectores de características (de naturaleza continua) siguiendo una adecuada ley probabilística: en este caso una *mixtura de densidades de Gaussianas* definidas de la forma siguiente:

$$B_j(x) = \sum_{l=1}^{L} c_{jl} N(x, \mu_{jl}, \sum_{jl})$$

donde:

- $b_j(x)$: función de distribución de densidad de probabilidad de emisión de vectores *x* de un HMM para el estado q_j.
- *L*: número de densidades de Gaussianas en la mixtura.

- c_{jl}: peso de la *l*-ésima Gaussiana de la mixtura del estado q_j . Debe cumplirse:

$$\sum_{l=1}^{L} c_{jl} = 1$$

- $N(x, \mu_{jl}, \sum_{jl})$: *l*-ésima componente de densidad de distribución Gaussiana en la mixtura del estado q_j , de media μ_{jl} y covarianza $\sum_{jl}$.

El número adecuado de densidades de probabilidad *L* en una mixtura depende, entre otros varios factores, de la variabilidad vertical que se espera en cada estado.

En cambio, el número adecuado de estados para modelar un cierto carácter o conjunto de caracteres, recae sobre la variabilidad horizontal. Ambos valores suelen determinarse empíricamente en cada tarea. También la cantidad de muestras disponibles para el entrenamiento de los modelos impone el número adecuado de parámetros de éstos (número de estados y Gaussianas).

Teniendo en cuenta lo anterior, los HMM tienen dos procesos estocásticos implícitos: uno describe las partes constituyentes del objeto (carácter) y otro las variaciones de dichas partes. Para el caso de OCR resulta adecuado utilizar HMMs con topología *izquierda derecha* acorde para el procesamiento de una secuencia de vectores de característica en función de su desplazamiento horizontal. Estos modelos se definen de forma tal que dados dos estados $q_i, q_j \in Q$ del HMM, es posible una transición entre ellos solo si $j \geq i$. Esto tiene, además, la ventaja deseable que debe aprenderse un número reducido de transiciones en la etapa de entrenamiento.

3 MARCO PRÁCTICO

3.1 DESCRIPCION INFORMAL DEL MODELO

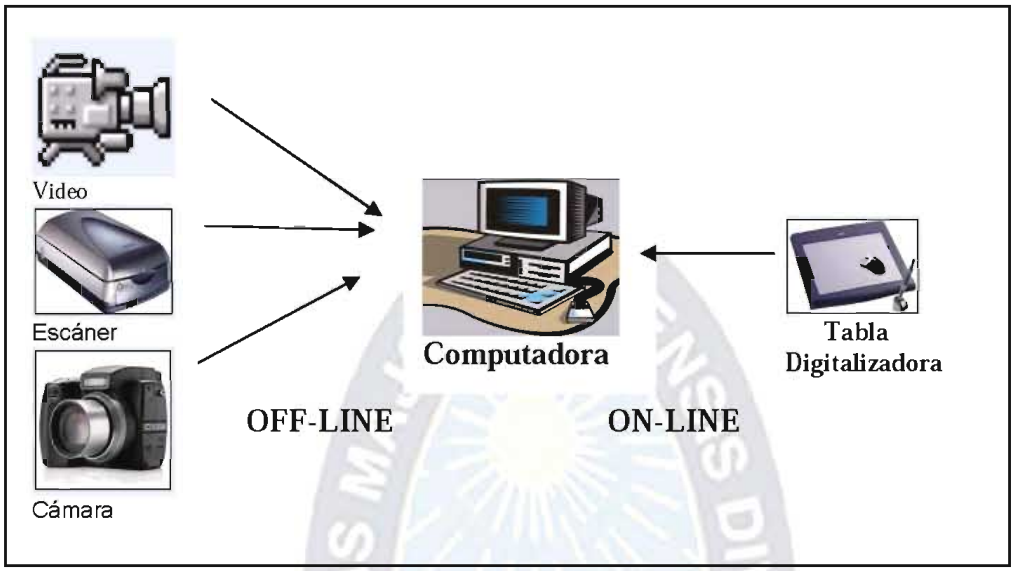
Para el reconocimiento de texto manuscrito continuo en lengua aymara, en la implementación del OCR (Reconocimiento Óptico de caracteres), que se ha explicado detalladamente en el capitulo 2, la implementación del reconocimiento de texto manuscrito continuo hasta llevarlos al lenguaje de maquina y su posterior reconocimiento de texto como resultado final, se plantean 2 formas de obtención de datos, como entrada de datos a un ordenador.

- On-Line: Los datos se obtienen en tiempo real mientras se escribe.
- Off-Line: Los datos son obtenidos por medio de escáneres, cámaras, todas estas en forma de imagen.

En el reconocimiento "On-line" el escritor esta conectado directamente por medio de un bolígrafo electrónico, lápiz óptico o dispositivos similares a un computador, esta escritura es registrado en función del tiempo y para el reconocimiento Off-line son por medio de video scanner, etc.

Para cumplir el propósito de esta investigación, consideraremos la entrada de datos la forma Off-line, introducción de datos mediante scanner como una imagen cualesquiera, para luego hacer un estudio de la misma cumpliendo los objetivos específicos planteados en el capitulo 1 de reconocimiento de texto manuscrito continuo en lengua aymara, ver figura 3.1 que describe la obtención de los datos de entrada desde sus 2 perspectiva, Off-line y On-line.

Figura 3.1 Obtención de Datos de entrada

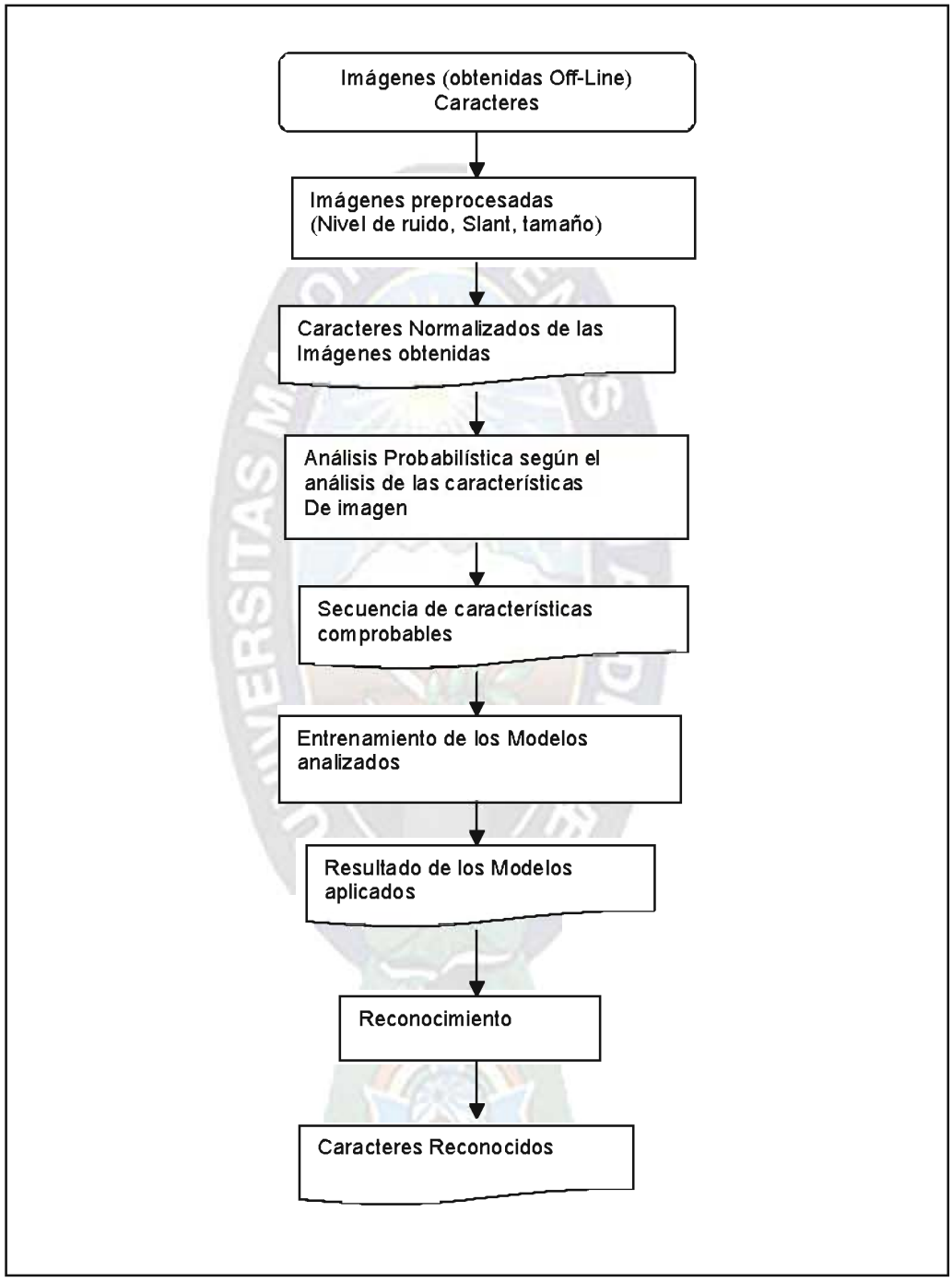


Fuente: [Elaboración Propia]

Para el reconocimiento Óptico de Caracteres (OCR) y su implementación se presenta el siguiente algoritmo de proceso, que plantea todos los pasos hasta el reconocimiento del texto, el preproceso de características aislados, buscando la reducción de ruido, extracción de características donde las imágenes de caracteres manuscritas son convertidas en vectores de características, entrenamiento de los modelos ocultos de Markov (HMMs) para determinar los valores de los parámetros de los modelos mediante aprendizaje y el reconocimiento con los vectores de características para definir a que carácter corresponde.

Detallaremos minuciosamente este proceso mediante un algoritmo ver figura 3.2.

Figura 3.2 Algoritmo de proceso del reconocimiento Óptico de Caracteres (OCR).



Fuente: [Elaboración Propia]

Tabla 3.1 Secuencia de pasos detalladas del algoritmo planteado

Procesos	Descripción
Imagen de Caracteres	Imágenes capturadas mediante Off-Line ya sea por video, escáner o cámara fotográfica para el recogimiento de texto manuscrito continuo
Imagen Preprocesada	Se hace una normalización de los diferentes tipos de escrituras manuscritas, en tamaño, inclinación y si existe un nivel de ruido en la imagen capturada en proceso.
Análisis de las características de la imagen	En el análisis de la imagen por ejemplo se verificara el nivel de gris en la imagen manejas en una matriz mxn.
Resultados Probabilísticas	Para la estimación de probabilidades se utilizaran los HMM que son maquinas de estados finitos.
Entrenamiento de modelos HMMs	Aplicando algoritmos para este propósito como Algoritmo "Backward, Forward", Algoritmo "Viterbi" para el entrenamiento adecuado de los parámetros.
Reconocimiento	Una vez entrenados adecuadamente las características, solo quedaría el reconocimiento del texto.
Caracteres reconocidas	Caracteres reconocidas de la extracción de datos

Fuente: [Elaboración Propia]

En el proceso en la secuencia del algoritmo los componentes que interactúan en este modelo del reconocimiento de texto manuscrito continuo mencionamos los Modelos Ocultos de Markov, algoritmos Backward Forward y Viterbi para un entrenamiento de los parámetros adecuadamente, el estudio de los atributos propios de la escritura manuscrita en lengua aymara ya que este idioma como se menciono anteriormente tiene su particularidad propia al ser escrita, en la escritura se puede analizar el tamaño, inclinación, el grosor o el nivel de gris con la que esta escrito manuscritamente el texto a ser analizado.

Las variables estudiadas en este modelo como el reconocimiento del texto manuscrito y los modelos estadísticos señaladas anteriormente como los Modelos de markov de capa oculta, cumplen un papel de vital importancia en el modelo para el entrenamiento de las

características o caracteres en análisis y el entrenamiento que debe cumplirse para el proceso de reconocimiento del texto.

3.2 RECONOCIMIENTO DE TEXTO MANUSCRITO CONTINUO

El reconocimiento de texto manuscrito continuo off-Line tiene que tener un alto grado de segmentación esto significa la compresión global de una frase escrita en aymara por ejemplo, en la vida real los individuos hacemos reconocimiento de manera natural para entender un texto manuscrito con un alto grado de segmentación, para que este texto sea descifrado o reconocido ya hemos aplicado de manera interna lo que se llama la segmentación pero de manera natural.

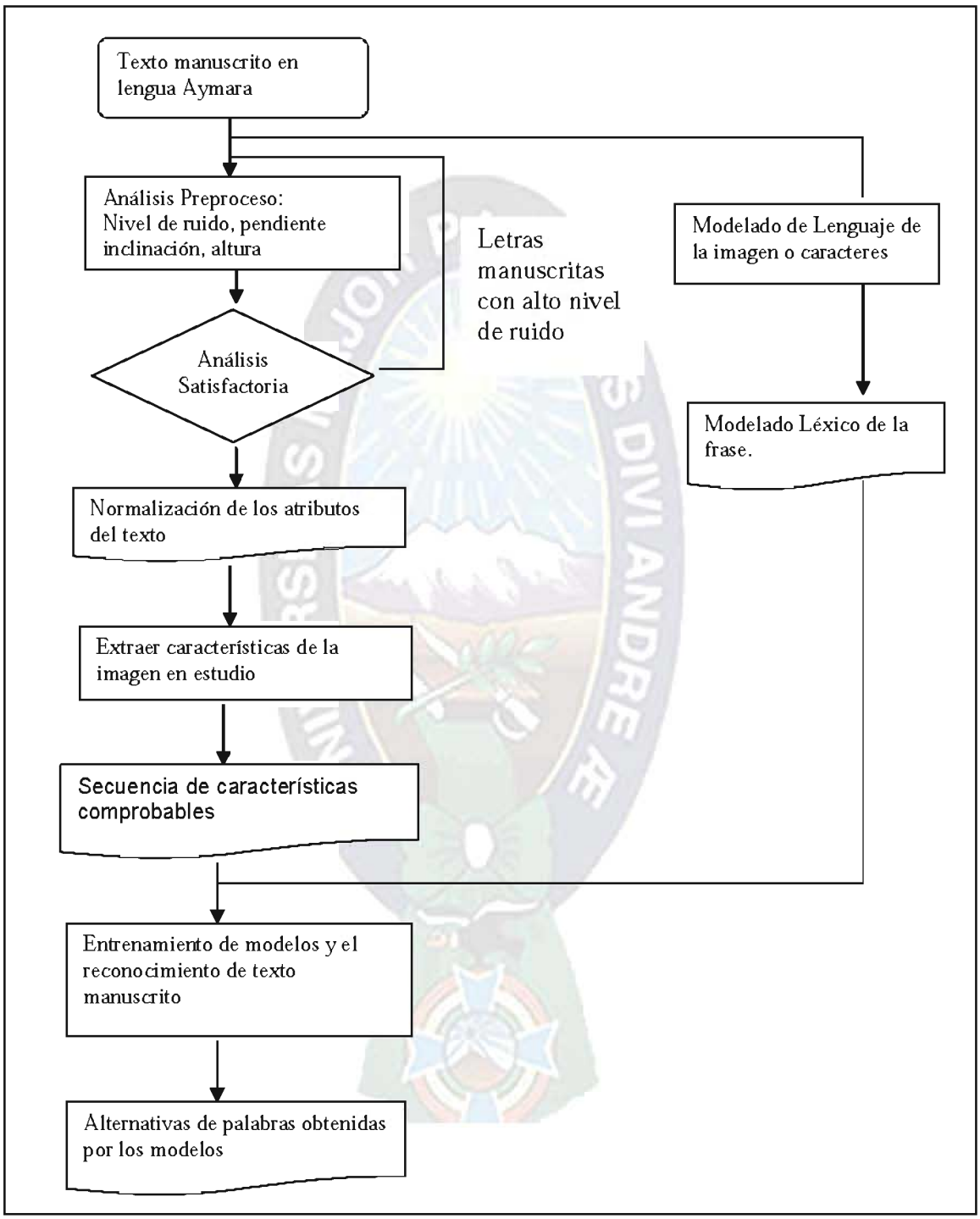
Para lograr la segmentación de la escritura manuscrita se considera básicamente adecuarse a un modelo simple, formulación de procesos de reconocimiento, utilizar técnicas adecuadas para asociar a los modelos planteados en los objetivos.

Para nuestro propósito llevaremos al estudio una frase escrita en aymara, consideraremos una cantidad numérica para el análisis, en lengua español por ejemplo este estudio podría ser la escritura en un cheque bancario.

Para el análisis de este propósito, haremos el estudio de la segmentación del texto, el estudio de las características propias, entrenamiento de los modelos, el modelado del lenguaje para la secuencia de palabras y el reconocimiento donde analizadas las características de la frase escrita en aymara luego tener palabras reconocidas.

El proceso planteado considerando los diferentes etapas desde la obtención de la imagen mediante off-line hasta obtener una secuencia de palabras o frases para el reconocimiento del texto, detallaremos en el algoritmo de la Figura 3.3.

Figura 3.3 Se observa el procedimiento del reconocimiento de texto.



Fuente: [Elaboración Propia]

Como es de nuestro conocimiento el escribir manuscritamente, trae algunos problemas como la escritura propia de cada individuo, por ejemplo se tiene escrituras poco legibles, escrituras en papeles donde los márgenes o cuadriculas muy resaltadas, no existe una normalización del nivel de base en la escritura, para el análisis de estos elementos por ejemplo tenemos el estudio de preproceso el nivel de ruido, pendientes de la escritura, inclinación, altura, anchura y el grosor del instrumento con la cual fue escrito.

3.3 PREPROCESO

3.3.1 NIVEL DE RUIDO: Como se puede observar en la figura 3.4, el nivel de ruido presentado es por ejemplo las líneas que se observan que no forman parte para el proceso del reconocimiento:

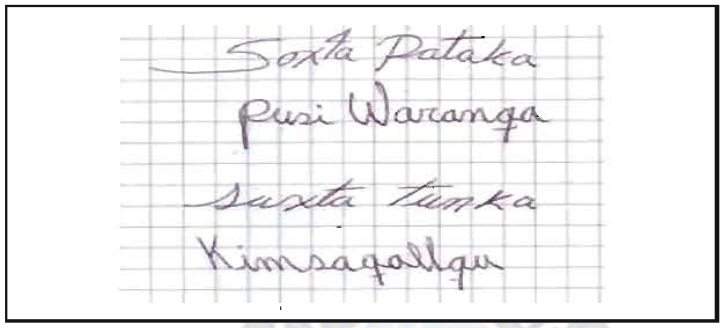
Figura 3.4 Datos con nivel de ruido

Fuente: [Elaboración Propia]

3.3.2 CORRECCION DE LA LINEA DE BASE

En este estudio tenemos por ejemplo en la figura 3.5, donde se puede observar las diferentes escrituras de distintos autores, en la cual estos escritos necesariamente debemos normalizarlos, esto según al eje x de la imagen considerando el eje vertical para Lugo hacer el proceso.

Figura 3.5 Análisis de Línea de Base

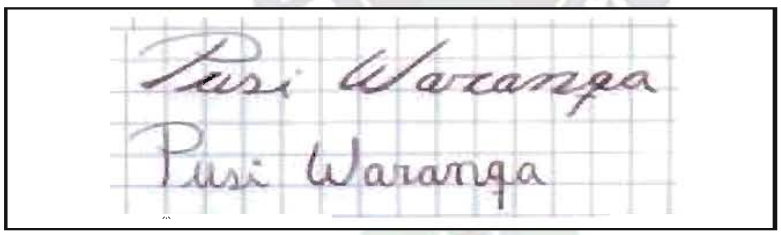


Fuente: [Elaboración Propia]

3.3.3 CORRECCION DE INCLINACION VERTICAL (Slant)

La corrección de la inclinación vertical de la escritura se debe hacer un suavizado de la misma con los modelos planteados, para hacer la normalización de los atributos propios de cada individuo con respecto a la línea vertical, como se observa en la figura 3.6.

Figura 3.6 Inclinaciones de escritura

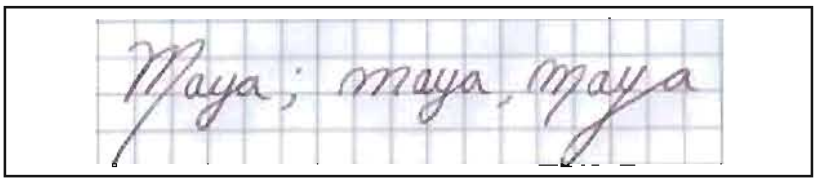


Fuente: [Elaboración Propia]

3.3.4 CONSIDERACIONES DE ALTURA

En la normalización de altura como se menciono anteriormente sobre los diferentes formas de escribir por cada autor se debe considerar, estas variante para poder estandarizar en formatos adecuados, observar figura 3.7.

Figura 3.7 Altura en las escrituras



Fuente [Elaboración Propia]

3.4 COMPONENTES

Los componentes o los modelos considerados, modelo markov de capa oculta (HMMs) y el modelado de lenguaje léxico, en el algoritmo de procesos, analizaremos como se dijo anteriormente una frase numérica como se observa en la figura 3.8, escrita en el idioma aymara, en el idioma español es mil doscientos diez.

Figura. 3.8 Palabra o frase en aymara (numérico)

Fuente: [Elaboración Propia]

Aplicando todos los procesos mencionados se consideran luego del preproceso una extracción de características de la imagen en estudio.

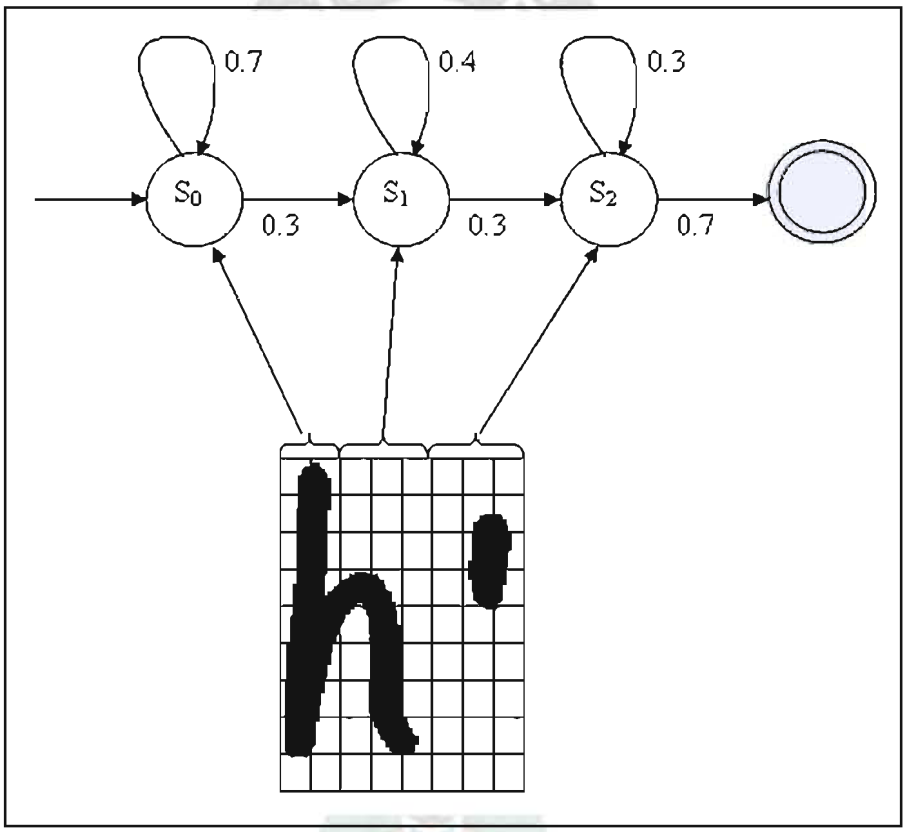
3.4.1 ANALISIS DE LAS CARACTERISTICAS

Luego del análisis del preproceso con todos los elementos considerados se extraen características para luego llevarlos estos en vectores de características en dimensiones determinadas o fijas, puede decirse también que estas características son fácilmente computables en imágenes binarias o lenguaje del computador.

Analizaremos un carácter en particular para el entrenamiento de la misma, con los Modelos ocultos de Markov (HMMs).

Se toma el contorno en estudio, luego de haber filtrado la imagen, binarizado, se clasifica esta matriz de la imagen en celdas de distribución simétrica de tal forma que podamos modelar con los modelos de markov de capa oculta (HMM) como se muestra en al figura 3.9

Figura 3.9 Muestra el modelado de un HMMs de las características h', que conforman los atributos propios del idioma aymara.

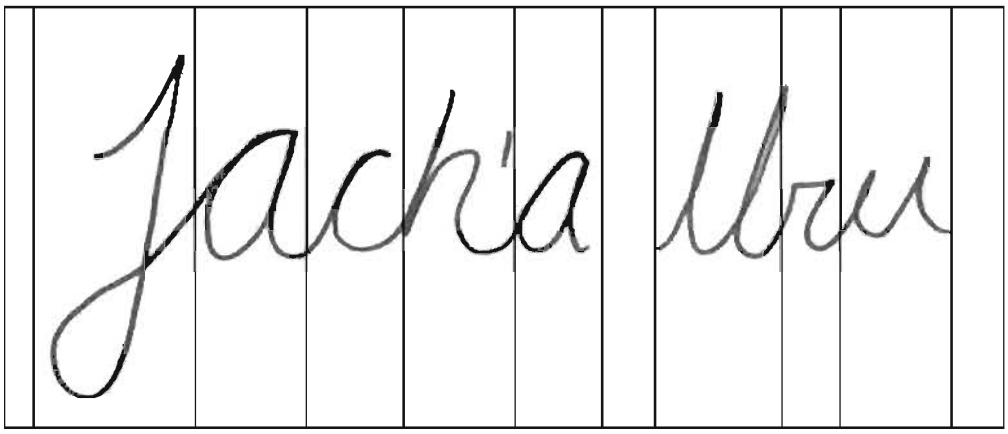


Fuente: [Elaboración Propia]

Se puede observar las probabilidades según el estado que modela, en este caso los estados para modelar son el número de columnas.

Para el estudio de frases se obtendrá un vector de características para que en ellos podamos extraer características independientes para luego analizarlos, modelarlos y hacer el respectivo entrenamiento del mismo veamos la figura 3.10 donde se toma la clasificación independiente de caracteres.

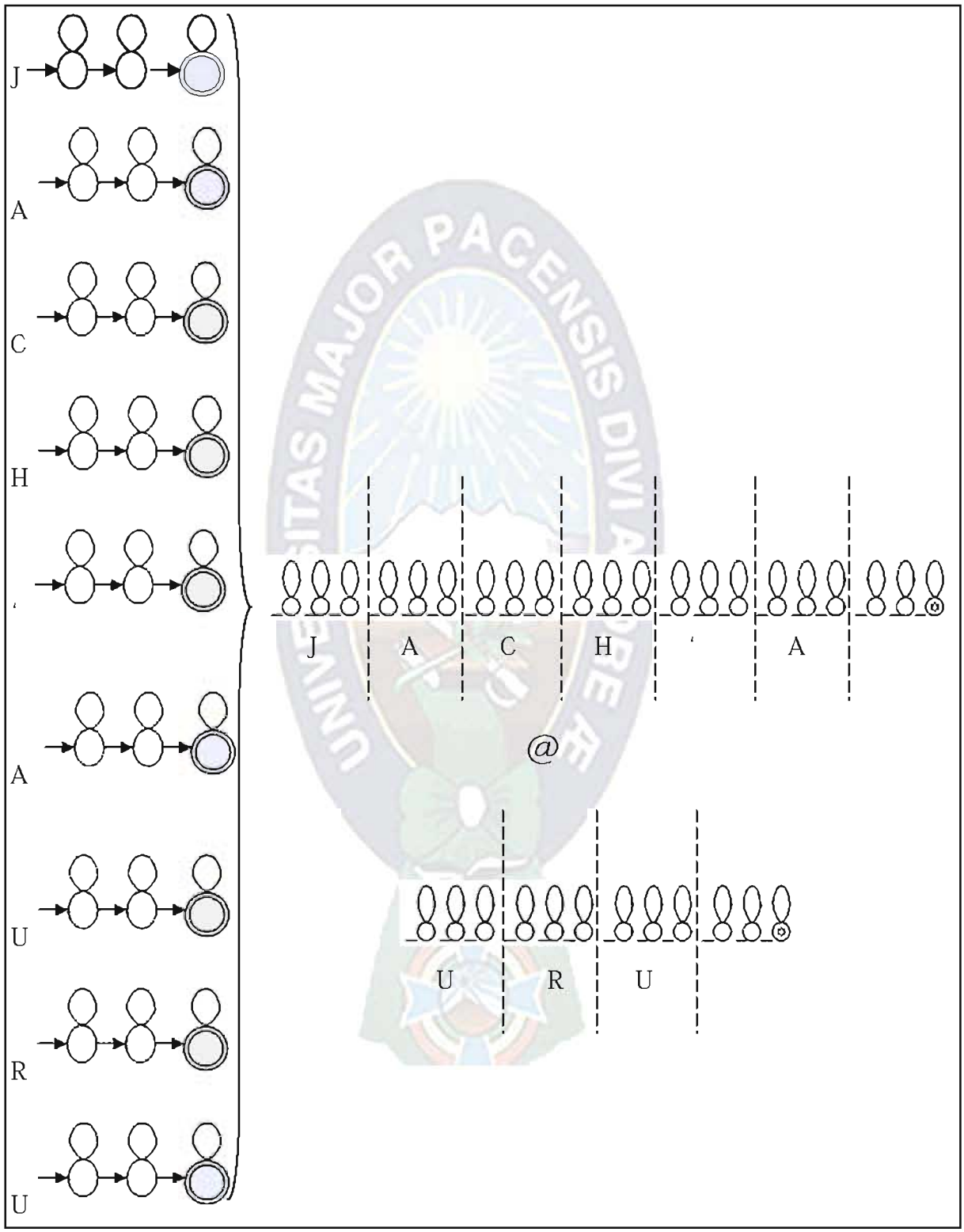
Figura 3.10 Clasificación de características independientes.



Fuente: [Elaboración Propia]

La extracción de características es importante para transformar la imagen preprocesada en una secuencia de vectores de características de dimensión fija, para obtener esto la imagen se divide en rejilla de celdas cuadradas homogéneas, en la figura 3.10 mostramos como se debe procesar la imagen por características separadas luego ser modeladas de manera independientes como se muestra en la figura 3.11

Figura 3.11 Modelado de las palabras “jach’a uru” .



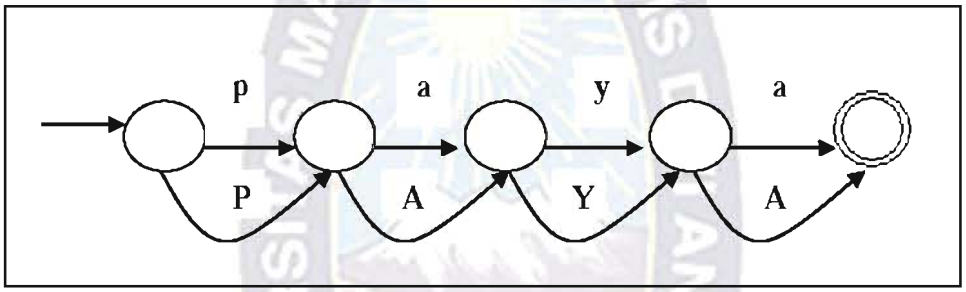
Fuente: [Elaboración Propia]

3.5 MODELADO LEXICO

En el modelado léxico se consideran las cadenas de caracteres o concatenación de caracteres que forma una escritura, considerando el Grafemario Aymará de 26 consonantes y 3 vocales.

Se consideran los modelos de Autómatas Estocásticos de Estados finitos (AEEF), esto para realizar la concatenación de la frase en estudio como se muestra en la Figura 3.12.

Figura 3.12: Autómatas estocásticos que modela la palabra Dos en Aymara

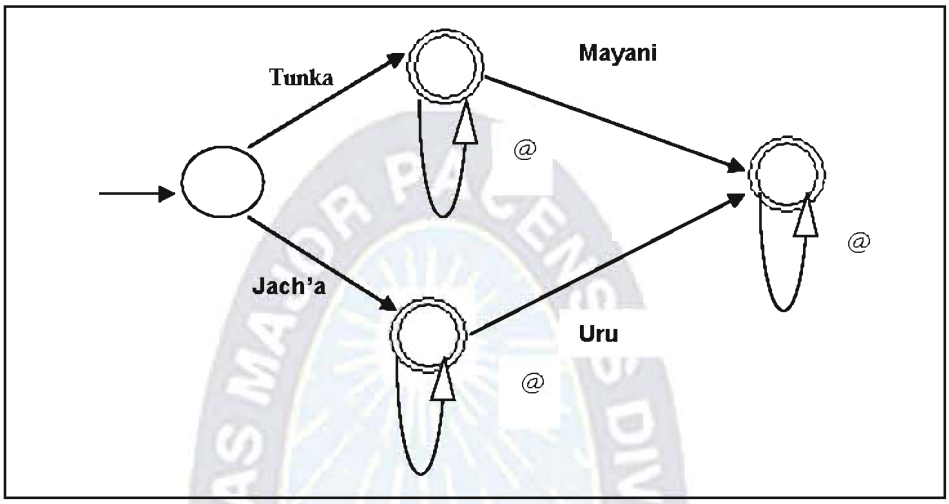


Fuente: [Elaboración Propia]

3.5.1 MODELADO DE FRASES EN AYMARA

En este estudio se modela frases o sea las concatenación no como se trato los caracteres si no se tiene frases, se toma como cadenas considerando las frases, se debe hacer también un estudio sobre el espacio entre palabras que existe, para ello utilizaremos el carácter @ (espacio entre caracteres) en el modelado de lenguaje (ML), ver figura 3.13.

Figura 3.13: Autómatas estocásticos que modela la frase tunka, tunka mayani, Jach'a, jach'a uru. En español díez, once, grande, grande día o (gran día).



Fuente: [Elaboración Propia]

3.6 MODELO DE LENGUAJE

Un modelo de lenguaje es un mecanismo para definir la estructura del lenguaje, es decir, para restringir adecuadamente las secuencias de unidades lingüísticas más probables.
En general son útiles en aplicaciones que exhiban una sintaxis y/o semántica completa, un buen modelo de lenguaje solamente debería aceptar frases correctas y rechazar aquellas secuencias de palabras incorrectas.

3.6.1 MODELADO DE LENGUAJE DE N-GRAMAS

Sirven de base tanto para el modelo de lenguaje empleado en reconocimiento de frases como para el sistema de clasificación de frases reconocidas. También se puede abordar lo que son los suavizados de N-gramas en el entrenamiento.

4 PROCESO DE INVESTIGACIÓN

4.1 DESCRIPCIÓN FORMAL

En el presente capítulo se presenta el reconocimiento de texto manuscrito continuo en un marco de una aplicación real, donde se consideran restricciones en estilo de escritura, tipo de letra y dirección de trazos.

En esta ocasión un nuevo requerimiento de la aplicación entra a formar parte en el sistema: el de la interpretación del texto manuscrito reconocido.

En otras palabras, la aplicación real de lo que se menciona, consiste esencialmente en clasificar frases en un conjunto reducido de categorías previamente establecidas.

Una buena parte de las frases manuscritas introducidas tienden a una amplia variedad de dificultades. Entre ellos, se destacan la aparición de letras mayúsculas y minúsculas dentro de una misma frase, apóstrofos escritas inadecuadamente, cambio de estilos en el trazo de la escritura, errores de ortografía, palabras tachadas, uso de abreviaturas (no estándares) y de símbolos para representar palabras, respuestas en idiomas diferentes, etc. Éstas provocan, en menor o mayor medida, la introducción de una mayor complejidad en el proceso de implementación del sistema, para automatizar el reconocimiento de dichas frases.

Se considera aquí la idea de los 3 diferentes niveles de conocimiento (morfológico, léxico y sintáctico).

Como se ha explicado, la inter-cooperación entre ellos permite a un sistema de reconocimiento de frases manuscritas en aymara en forma efectiva la tarea de decodificación

de secuencias de vectores de características (provenientes de su entrada) y producir secuencias de palabras reconocidas (en su salida).

Como los requerimientos de la nueva aplicación exigen además una clasificación de las frases reconocidas, añadimos un nuevo nivel de conocimiento al sistema: el semántico.

Este nivel se integrará con los otros 3 niveles y será responsable de la correcta interpretación de frases reconocidas, categorizando cada una de ellas, asignándoles un significado.

Nuevamente, para el modelado de estos 4 niveles de conocimiento recurrimos a las máquinas de estados finitos que resultan muy convenientes, permite integrar todos ellos entre sí de una manera natural y sencilla.

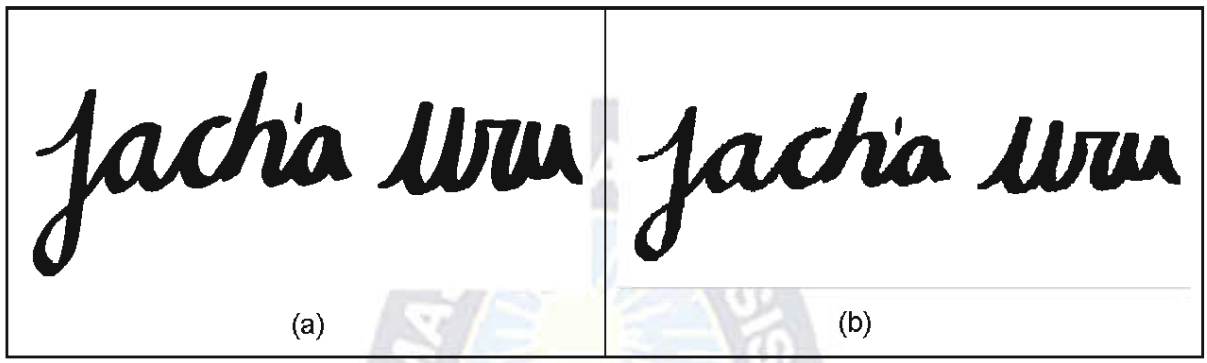
El nivel morfológico es modelado utilizando HMMs (para cada uno de los caracteres); el nivel léxico es modelado por autómatas de estados finitos (para cada una de las palabras); y finalmente, el nivel sintáctico y semántico son modelados en este caso por *N*-gramas, representados también como máquinas de estados finitos.

Para finalizar, el objetivo final de este capítulo consiste en diseñar un sistema que automatice lo mejor posible el proceso de reconocimiento de la escritura en lengua aymara y clasificación (interpretación) de las respuestas manuscritas, disminuyendo al máximo la intervención humana en dicha tarea.

4.2 DESCRIPCIÓN DE PROTOTIPO

En el análisis de escritura manuscrita, consideraremos imágenes escaneados en papel blanco con escritura de color negro u otro, de manera que se captura una imagen en mapa de bits (.bmp), luego de ello convertimos la imagen original en imagen binaria, para luego hacer una segmentación de imagen binarizada, hacemos el análisis de las características individuales con los modelos estadísticos planteados, En la captura de una imagen original en Matlab considera los diferentes tipos de lectura de imágenes, siempre tomando como una matriz bidimensional o tridimensional según el caso en estudio, como ser los .bmp, .jpeg, .jpg, .tif, .tiff, .pnm, etc. En nuestro caso trabajaremos como imágenes de mapa de bits (.bmp). veamos tratamiento de imágenes en figura 4.1 y Figura 4.2.

Figura 4.1 Imagen de escritura con un trazo grueso. (a)Imagen manuscrita original obtenida mediante un escáner, (b) Imagen binarizada en ceros y unos, los ceros son oscuros y los unos son blancos.



(a) (b)

Fuente: [Elaboración Propia]

Figura 4.2 Imagen de escritura con trazo delgado (Lápiz delgada), (c) Imagen original obtenida mediante un escáner, (d) Imagen binarizada en ceros y unos.

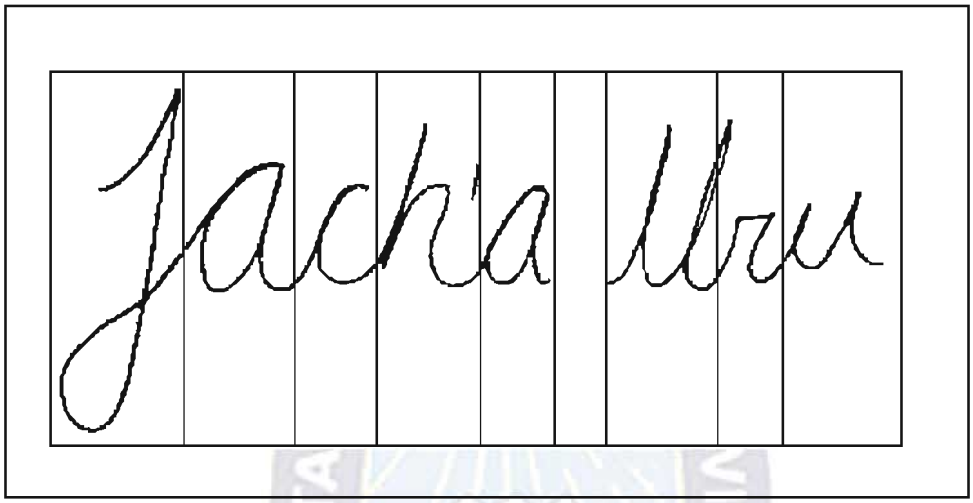
(c) (d)

Fuente: [Elaboración Propia]

4.2.1 EXTRACCIÓN DE CARACTERÍSTICAS

Lego de Binarizar la imagen manuscrita se hace el análisis de la extracción de características individuales, ya una vez binarizada procesamos en manejo de píxeles de la imagen, la extracción de características como se menciono anteriormente la obtendremos en un vector de características, ver Figura 4.3

Figura 4.3 Separación de la escritura manuscrita en caracteres individuales para implementar estas características con los modelos estadísticos planteados y así concluir en el reconocimiento del carácter final en lenguaje del computador (ASCII).



Fuente: [Elaboración Propia]

4.3 PROGRAMA

En el proceso de reconocimiento de texto manuscrito continuo, primeramente haremos un estudio minucioso de un carácter en concreto como modelamos este carácter en sus diferentes etapas hasta llegar al reconocimiento del texto manuscrito en el idioma aymara.

Para implementar el prototipo se utilizara Matlab 7.0 por sus facilidades o comandos implementados en este software en el campo de la matemática

Algoritmo 1 Pseudocódigo que implementa todo el proceso de reconocimiento llamando funciones.

```
clear all;
  I=imread('Imagen.bmp');
info=imfinfo('Imagen.bmp');
if info.ColorType=='truecolor'
 I=rgb2gray(I);
 else I=I;
end
imshow(I)
I2=sobel(Id,R,C,2.5); %Halla gradiente de sobel
[f,c]=size(I);J=Binarizacion(I,f,c,I2);
Id=double(I); [R C]=size(I)
figure;imshow(I2)
```

```
umbral=UmbralOptimo(Id,R,C,I2); %Halla umbral optimo
VectorCaracteristica=Chmm_def(Id);
BD=base_datos()
```

Algoritmo 2. Pseudocódigo del algoritmo de binarizacion de una imagen

```
Algoritmo de Binarizacion.

function imagenBin=Binarizacion(imagen, filasImagen, columnasImagen, umbral)

for i=1:filasImagen
  for j=1:columnasImagen
    if imagen(i,j)>umbral
        imagenBin(i,j)=1;
    else
        imagenBin(i,j)=0;
    end
  end
end
```

Algoritmo 3 Filtrado de la imagen para el tratamiento del contorno de interés

```
function imagenFiltrada=Filtro(imagen, filasImagen, columnasImagen, kernel)

T=[0 0 1;0 1 0;1 0 0];
kernelRotado=T*kernel*T;

imagenFiltrada=zeros(filasImagen, columnasImagen);

for i=1+1:filasImagen-1
  for j=1+1:columnasImagen-1
    imagenFiltrada(i,j)=sum(sum(imagen(i-1:i+1,j-1:j+1).*kernelRotado));
  end
end
```

Algoritmo 4. Pseudocódigo que implementa Modelos Ocultos de Marcok (HMMs) en el análisis de características

```
function chmm_def(fhmm)
nc=10;ng=7;grupo=cell(ng,1);Np=zeros(ng,1);
for ig=1:ng
  agrupo{ig}=[1 2 3];
end
for ig=1:ng
  Np(ig)=length(agrupo{ig})-1;
end
```

```
Ngauss=cell(ng,1);
for ig=1:ng
  Ngauss{ig}=6.*ones(Np(ig),1);
end
men=1;salto=1;maxiter=10;umbral=0.05;maxitermi=5;
A=cell(nc,ng);B=cell(nc,ng);Med=cell(nc,ng);Var=cell(nc,ng);Pi=cell(nc,ng);
for ic=1:nc,
  for ig=1:ng
    A{ic,ig}=zeros(Ne(ic,ig),Ne(ic,ig));
    B{ic,ig}=cell(Np(ig),1);  Med{ic,ig}=cell(Np(ig),1);Var{ic,ig}=cell(Np(ig),1);
    for ip=1:Np(ig)
      B{ic,ig}{ip}=cell(Ne(ic,ig),1); Med{ic,ig}{ip}=cell(Ne(ic,ig),1);
      Var{ic,ig}{ip}=cell(Ne(ic,ig),1);
      for ie=1:Ne(ic,ig),
        B{ic,ig}{ip}{ie}=zeros(Ngauss{ig}(ip),1);
        Med{ic,ig}{ip}{ie}=zeros(Ngauss{ig}(ip),agrupo{ig}(ip+1)-agrupo{ig}(ip));
        Var{ic,ig}{ip}{ie}=zeros(Ngauss{ig}(ip),agrupo{ig}(ip+1)-agrupo{ig}(ip));
      end
    end
    Pi{ic,ig}=zeros(Ne(ic,ig),1);
  end
end
vTEST=[' salhmm']; salhmm=cell(nc,ng);
guardar=['save ',fhmm,vDB,vHMM,vTEST,' iniciar vDB vHMM vTEST guardar'];
eval([guardar]);
```

4.4 PRESENTACIÓN DEL MODELO

Para la implementación del interfaz de usuario, y la presentación del modelo del reconocimiento de texto manuscrito, se utilizara el VRU (Validación de Requisitos de Usuario) este método puede integrarse dentro de cualquier método de producción de software que utilice casos de uso para describir los requisitos funcionales y algunas variantes de los diagramas de secuencia para describir las interacciones internas dentro del sistema.

VRU esta conformado por un componente proceso, una arquitectura de modelos y una notación. El componente proceso describe mediante flujos de trabajo las distintas fases del ciclo de desarrollo que cubre VRU, desde la verificación de requisitos hasta la generación de las interfaces y la validación de los requisitos iniciales.

Los componentes principales del VRU como las fases del ciclo de vida, los flujos de trabajo las actividades y sus salidas se puede observar en la tabla 4.1.

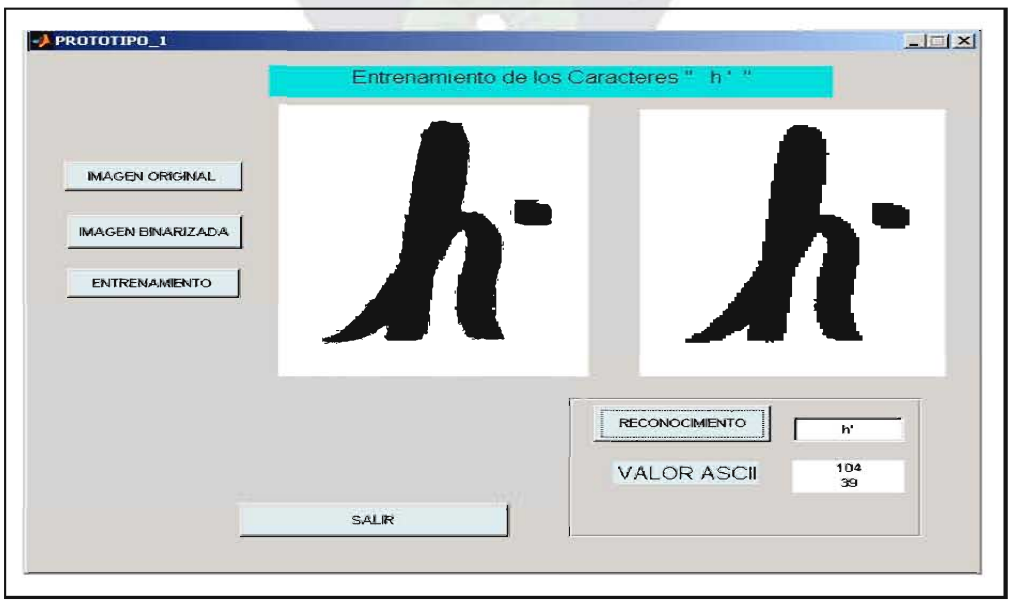
Tabla 4.1 Muestra el proceso VRU para una mejor interfaz de usuario

Etapa del ciclo de vida	Flujo de Trabajo	Actividades de VRU	Entregables
Análisis preliminar	Análisis Externo	Interiorizar Proyecto	Diccionario
			Requisitos Informales
		Crear perfil de usuario	Mapa de roles de usuario
		Elaborar ámbito Sistema	Modelo de dominio
Análisis detallado	Análisis Funcional	Crear modelo funcional	Actores
			Casos de Uso
			Servicios Funcionales
	Análisis Interno	Modelizacion de objetos	Diagrama de clases
		Síntesis de casos de uso	Diagrama MSC
	Generación	Generar interfaz	Modelo de presentación
		Generar dinámica	Diagramas de transición
Validación	Animación	Ejecución interfaz	

Esta presentación principal del prototipo se muestra el modelado de un carácter escrito manuscritamente en el idioma aymara (h'), ya que esto es el objetivo del estudio de las características propias del idioma aymara ver figura 4.4.

Caso 1

Figura 4.4 Prueba de Reconocimiento de los caracteres manuscritos " h ' "

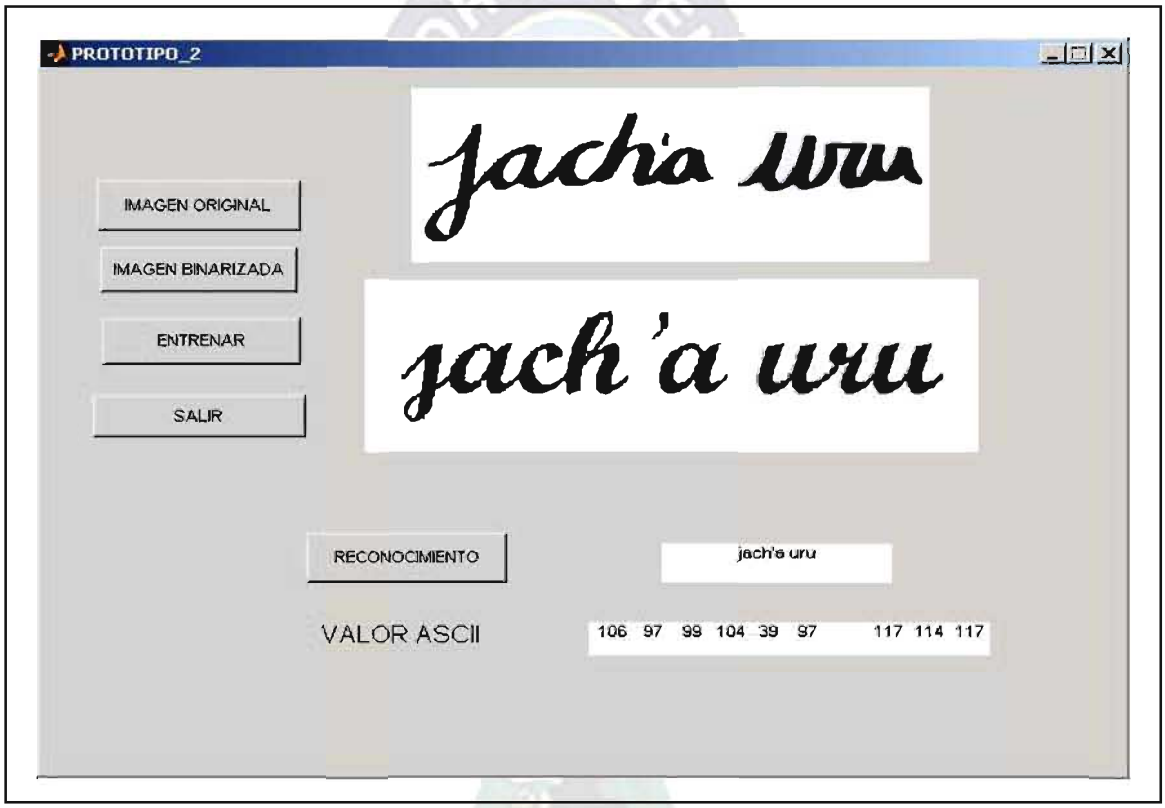


Fuente: [Elaboración Propia]

Caso 2

Análisis de la frase manuscrita en aymara "jach'a Uru", obtención de imagen escaneada, binarizacion de la misma, el entrenamiento con los modelos y el valor ASCII luego de haber reconocido la frase en estudio, ver figura 4.5.

Figura 4.5 Prueba de reconocimiento de las palabras manuscritas "jach'a uru"



Fuente: [Elaboración Propia]

Caso 3

Aplicando todo el procedimiento para el reconocimiento de texto manuscrito, como también el reconocimiento óptico de caracteres (OCR), como se menciono anteriormente, para el modelado de caracteres en el morfológico aplicamos HMMs, el nivel léxico es modelado por autómatas de estados finitos y finalmente para el modelado sintáctico utilizamos los N-gramas.

Se puede observar en la Figura 4.6 los resultados obtenidos son como esperamos sin fallos de reconocimiento y su valor ASCII correspondiente.

Figura 4.6 Prueba de reconocimiento de la palabra manuscrita "jutaskiwa"

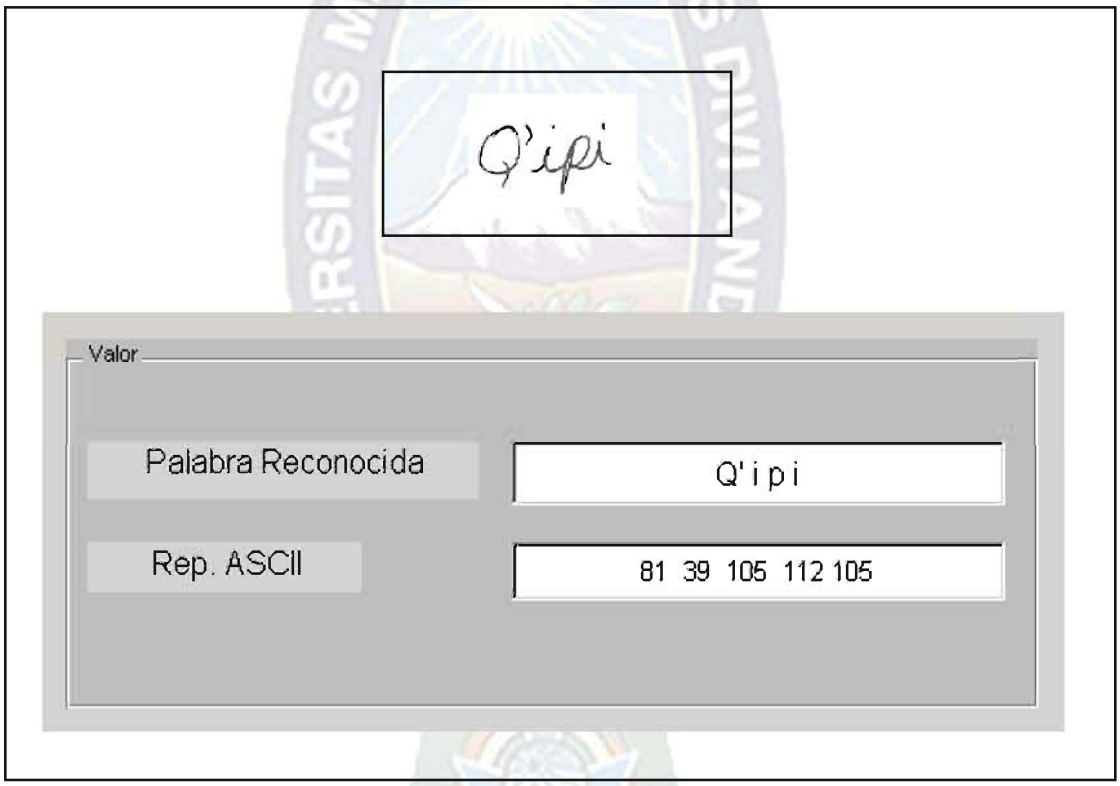
Fuente: [Elaboración Propia]

Caso 4

Aplicando todo el procedimiento para el reconocimiento de texto manuscrito, como también el reconocimiento óptico de caracteres (OCR), como se menciono anteriormente, para el modelado de caracteres en el morfológico aplicamos HMMs, el nivel léxico es modelado por autómatas de estados finitos y finalmente para el modelado sintáctico utilizamos los N-gramas.

Se puede observar en la Figura 4.7 los resultados obtenidos son como esperamos sin fallos de reconocimiento y su valor ASCII correspondiente.

Figura 4.7 Prueba de reconocimiento de la palabra manuscrita "Q' ipi"



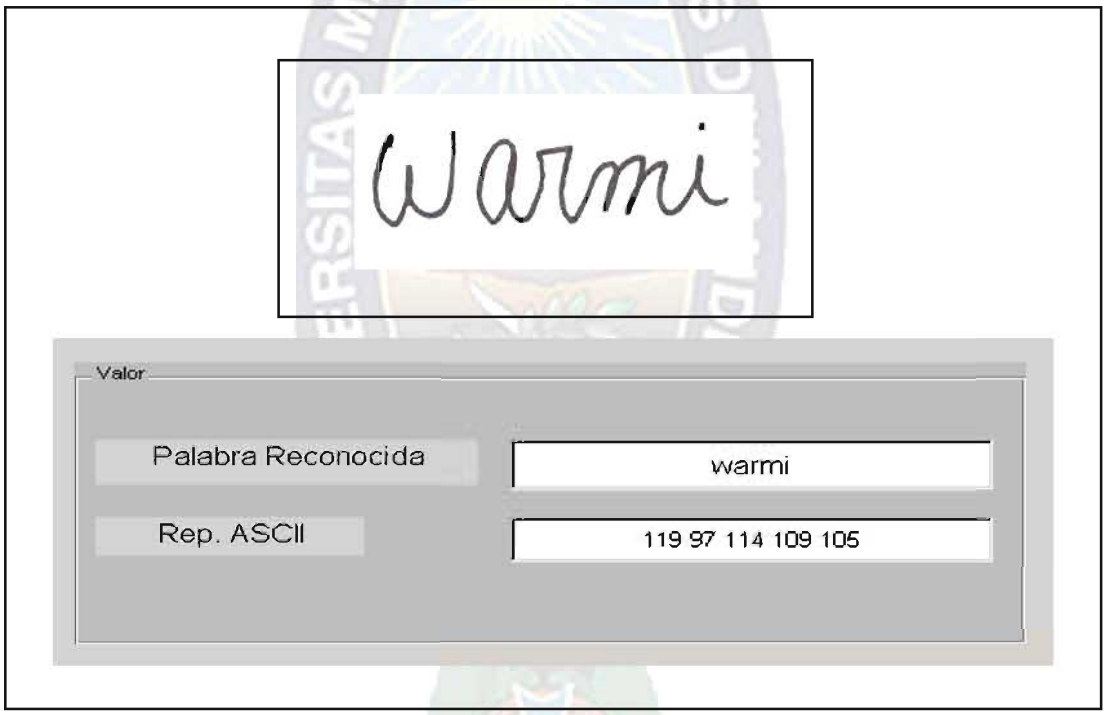
Fuente: [Elaboración Propia]

Caso 5

Aplicando todo el procedimiento para el reconocimiento de texto manuscrito, como también el reconocimiento óptico de caracteres (OCR), como se menciono anteriormente, para el modelado de caracteres en el morfológico aplicamos HMMs, el nivel léxico es modelado por autómatas de estados finitos y finalmente para el modelado sintáctico utilizamos los N-gramas.

Se puede observar en la Figura 4.8 los resultados obtenidos son como esperamos sin fallos de reconocimiento y su valor ASCII correspondiente.

Figura 4.8 Prueba de reconocimiento de la palabra manuscrita "warmi"

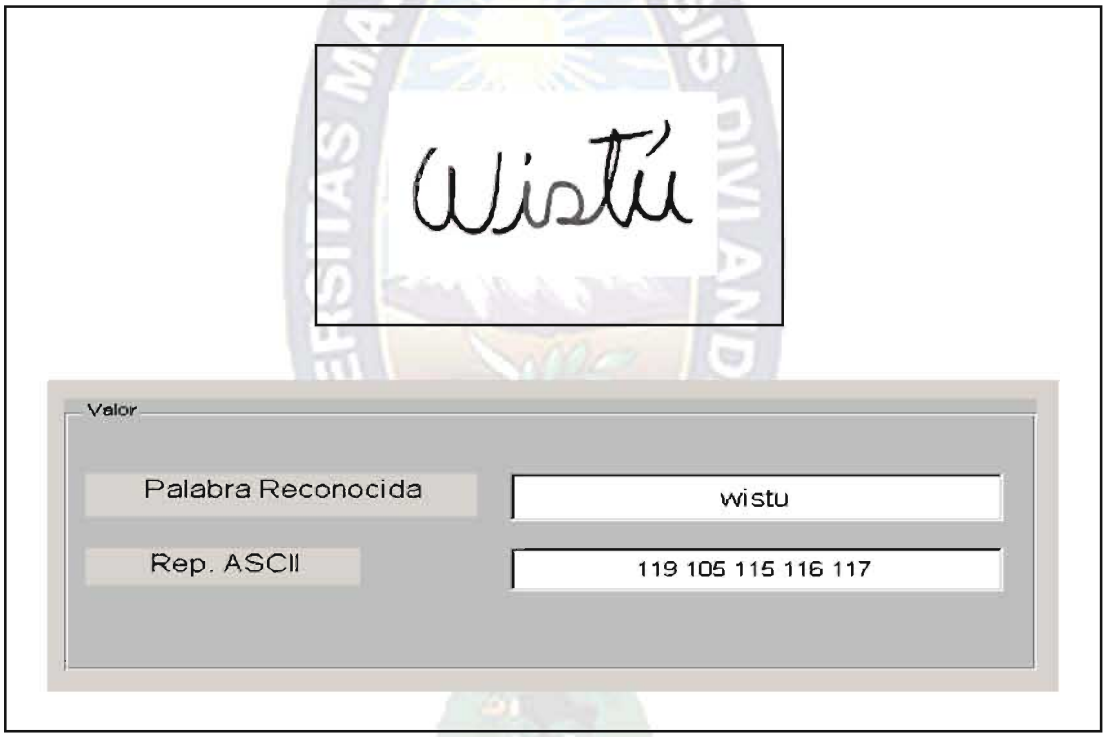


Fuente: [Elaboración Propia]

Caso 6

En este caso se puede observar los resultados en la Figura 4.9 no satisfactoria esto es debido a que la palabra manuscrita no cumple con las restricciones planteadas en la segmentación de la misma se tienes escritura a mano en el nivel ascendente encima de un carácter como son el carácter "u" y " ' ", donde no se hace el análisis de obtención de las características de manera adecuada, no reconoce el apostrofo como parte de la escritura en aymara como vera solo reconoce 5 caracteres.

Figura 4.9 Prueba de reconocimiento de la palabra manuscrita "wist'u"



Fuente: [Elaboración Propia]

Caso 7

De la misma forma se puede observar en la figura 4.10 el reconocimiento del texto manuscrito que no satisface el resultado esperado por lo que los caracteres "s, w" no están escritas adecuadamente, donde los modelos desconoce estas características y obtenemos resultados diferentes del esperado.

Figura 4.10 Reconocimiento de la palabra manuscrita "justaskiwa"

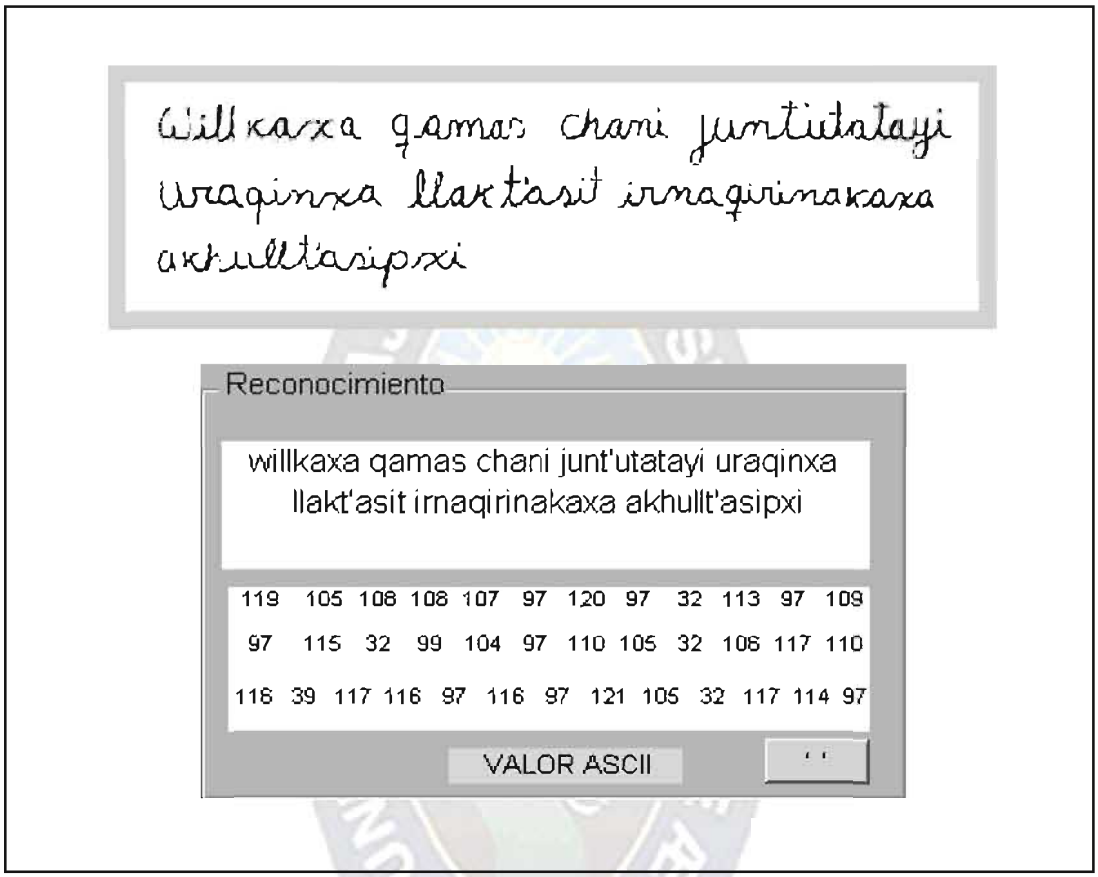
Fuente: [Elaboración Propia]

Caso 8

El nivel morfológico es modelado utilizando HMMs (para cada uno de los caracteres); el nivel léxico es modelado por autómatas de estados finitos (para cada una de las palabras); y finalmente, el nivel sintáctico y semántico son modelados en este caso por *N*-gramas, representados también como máquinas de estados finitos.

Ver figura 4.11 modelado de frases manuscritas

Figura 4.11 Reconocimiento de frase manuscrita “willkaxa qamas chani junt’utatayi uranqinxa llakt’asit irnaqirinakaxa akhullt’asipxi”.



Fuente: [Elaboración propia]

Caso 9

El nivel morfológico es modelado utilizando HMMs (para cada uno de los caracteres); el nivel léxico es modelado por autómatas de estados finitos (para cada una de las palabras); y finalmente, el nivel sintáctico y semántico son modelados en este caso por *N*-gramas, representados también como máquinas de estados finitos.

Ver figura 4.12 modelado de frases manuscritas sin restricción, escritura con pendiente de línea izquierda, derecha y escritura de usuario inadecuada.

Figura 4.12 Reconocimiento de frase manuscrita "wali uñakipat markax arsutama ist'aña munapxi kikip sarawinak q'ayachasina".

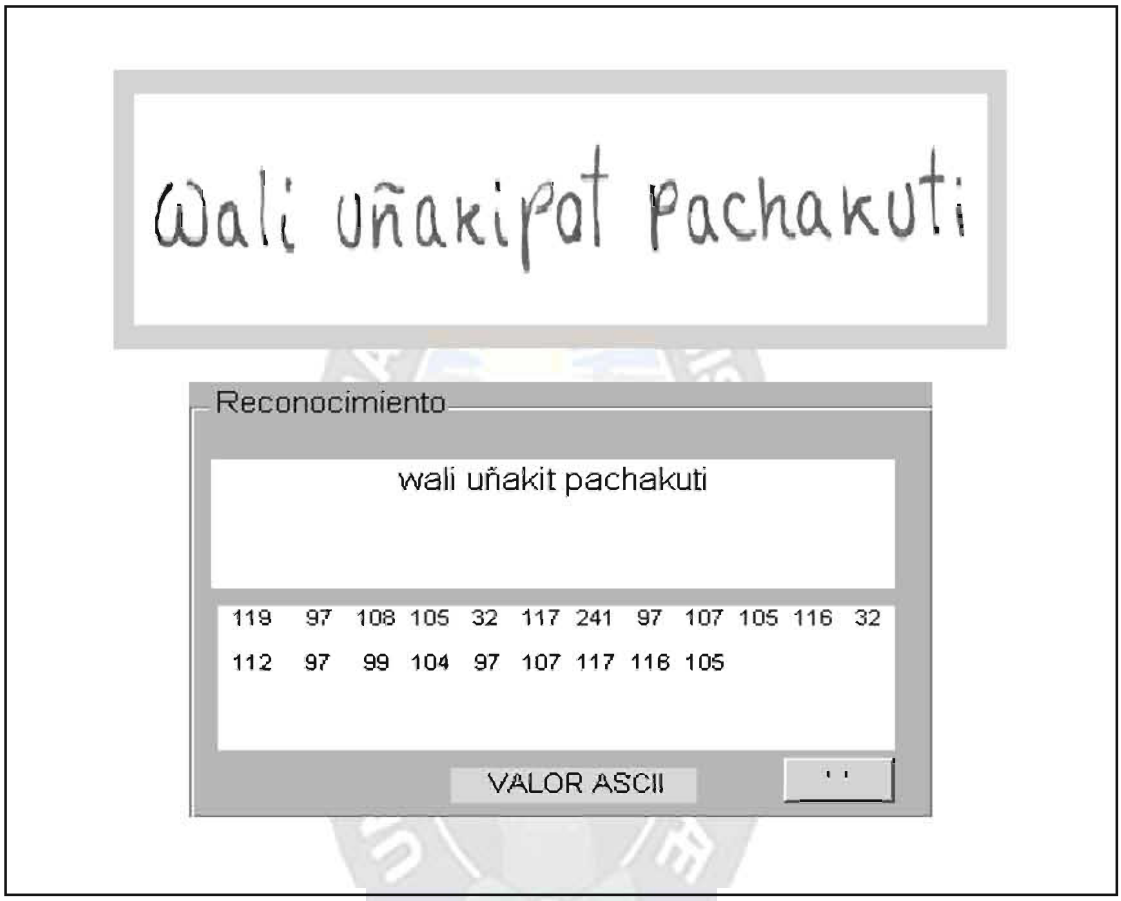
Fuente: [Elaboración Propia]

Caso 10

El nivel morfológico es modelado utilizando HMMs (para cada uno de los caracteres); el nivel léxico es modelado por autómatas de estados finitos (para cada una de las palabras); y finalmente, el nivel sintáctico y semántico son modelados en este caso por *N*-gramas, representados también como máquinas de estados finitos.

Ver figura 4.12 modelado de frases manuscritas que simula ser en imprenta sin restricción, Observe los resultados del modelo obtenido, no se reconoce el carácter p debido al filtrado de la imagen y el carácter "a" modela como el carácter "o" en idioma aymara este carácter no puede ser representado y el modelo no lo acepta por que no hay escritura en aymara con el carácter "o".

Figura 4.13 Reconocimiento de frase manuscrita que simula ser escrita en imprenta "wali uñakipat pachakuti".



Fuente: [Elaboración Propia]

4.5 ANÁLISIS DE DATOS Y RESULTADOS

Considerando estos datos probados con respecto a estas tecnologías, primeramente realizaremos el estudio de los resultados obtenidos con el modelo planteado en esta tesis, según los casos planteados.

Para el caso 1 Tenemos:
Numero de caracteres de entrada (NCE) o caracteres en imagen a considerar.
NCE=2
Numero de caracteres que reconoció el modelo (NCR)

NCR=2

Para calcular utilizaremos la precisión porcentual P que será igual al NCR/NCE multiplicado por 100 que nos da resultado en porcentaje de reconocimiento de las características en estudio.

Tabla 4.2 Análisis de Resultados

Caso 1 P=(NCR/NCE)*100= $\frac{2}{2}*100\,\%=100\%$	**Caso 2** NCE=10; NCR= 10; P=$\frac{10}{10}*100\,\%=100\%$;
Caso3 NCE=9; NCR=9; P=$\frac{9}{9}*100\ \%=100\%$	**Caso 4** NCE=5; NCR=5; P=$\frac{5}{5}*100\,\%=100\%$
Caso 5 NCE=5; NCR=5; P=$\frac{5}{5}*100\,\%=100\%$	**Caso 6** NCE=6; NCR=5; P=$\frac{5}{6}*100\,\%=83.333\%$
Caso 7 NCE=9; NCR=7; P=$\frac{7}{9}*100\,\%=77.778\%$	**Caso 8** NCE=81; NCR=81; P=$\frac{81}{81}*100\,\%=100\%$
Caso 9 NCE=82; NCR=34; P=$\frac{34}{82}*100\,\%=40.476\%$	**Caso 10** NCE=23; NCR=21; P=$\frac{21}{23}*100\,\%=91.304\%$

Fuente: [Elaboración Propia]

Ahora obtenemos el total de preedición porcentual $PT=\sum \frac{P}{N^{o} Casos}$

$$\mathbf{PT} = \frac{100+100+100+100+100+83.333+77.778+100+40.476+91.304}{10} = 89.289\%$$

4.5.1 RESULTADOS DEL ENTRENAMIENTO DE HMMS

Se hizo el entrenamiento con estados diferentes para el modelado de características, los resultados de muestran en la tabla, con diferentes numero de densidades Gaussianas.

Tabla 4.3 Error porcentual de clasificación usando HMMs con estados diferentes.

Numero de estados	Numero Gaussianas por estado						
	1	2	3	4	8	16	32
2	12.0	15.4	11.6	10.5	14.2	17.6	11.8
2.5	8.7	8.8	8.2	8.2	11.3	15.8	16.9
3	14.6	12.8	11.8	10.8	12.6	14.7	13.2

Fuente: [Elaboración Propia]

4.5.2 ANÁLISIS DE RESULTADO CON AUTÓMATAS ESTOCÁSTICOS DE ESTADOS FINITOS (AEEF)

Se tomaron en cuenta las restricciones de escritura manuscrita.

1. Modelado del lenguaje origina, con espacios en blanco por cada palabra y las frases respectivamente
2. Frases con sintaxis correctas
3. Tomando restricciones adecuadas para el modelado

Tabla 4.4 Análisis de Resultados

AEEF	Error Porcentual (%)
1	4.0
2	4.6
3	3.1

Fuente. [Elaboración Propia]

4.5.3 RESULTADOS DE RECONOCIMIENTO

En la tabla 4.5 se muestran los resultados en la fase de reconocimiento con valore Gaussianas con estados individuales en los HMMs, con unigramas y bigramas, no se consideran signos de puntuación, utilizando un total considerable de gramas presenta menor error porcentual.

Tabla 4.5 Resultados de la tasa de error de reconocimiento de palabras

	Gaussianas			
	2	4	8	16
FR	5.3%	48.3%	47.5%	39.8%

	1-Grama	2-Grama
FR	48.6	44.0

Fuente: [Elaboración Propia]

4.5.4 ESTUDIO DE HIPOTESIS

El reconocimiento de texto manuscrito continuo a través de la utilización de los Modelos Ocultos de Markov (HMMs) permite un nivel de fiabilidad del ochenta y cinco por ciento, con relación a otras aplicaciones en el reconocimiento automático de texto manuscrito continuo.

Para el análisis y prueba de esta hipótesis, conjuntamente con los resultados obtenidos del modelo, primeramente utilizaremos datos de tesis realizados en nuestra carrera con relación a reconocimiento de patrones que utilizan otra tecnologías, se tiene las siguientes tesis y la hipótesis plantadas en estas tecnologías.

Tesis: Reconocimiento de Escritura a Mano mediante Redes Neuronales

De manera textual la hipótesis señala: Construir un modelo reconocedor de Patrones digitales, sobre bases teóricas de las redes neuronales artificiales, que permita el reconocimiento de caracteres manuscritos [Aliaga T.; 2002].

Tesis: Reconocimiento Automático de Caracteres Manuscritos Continuos con Redes Neuronales y Lógica Difusa.

De manera textual la hipótesis señala: Diseñar un programa de mayor precisión en el reconocimiento de de caracteres manuscritos, empleando los paradigmas de redes neuronales artificiales y lógica difusa, con respecto de otro programa basado solo en el paradigma de redes neuronales [Alvarado R.; 2006].

Una prueba de hipótesis es un procedimiento basado en la evidencia muestral y en la teoría de probabilidad que se emplea para determinar si la hipótesis es un enunciado razonable y no debe rechazarse o si no es razonable y debe ser rechazado, sobre la hipótesis planteada en esta tesis consideraremos para demostrar lo que son la hipótesis nula y la hipótesis alternativa, luego seleccionaremos el nivel de significancia.

Hipótesis nula H_0: Se afirma el valor del 85% de efectividad

Hipótesis alterna H_1: Afirmación que se aceptara si los datos muestrales proporcionan evidencia de que la hipótesis nula es falsa.

Calculo de la desviación estándar

$$VP=\sqrt{\frac{\sum (P_i - PT)^2}{n}}$$

Tabla 4.6 Análisis de Resultados

P_i	$(P_i\text{-}PT)^2$	
100	$(100\text{-}89.3)^2$	= 114.5
100	$(100\text{-}89.3)^2$	= 114.5
100	$(100\text{-}89.3)^2$	= 114.5
100	$(100\text{-}89.3)^2$	= 114.5
100	$(100\text{-}89.3)^2$	= 114.5
83.33	$(83.33\text{-}89.3)^2$	= 123.43
77.78	$(77.78\text{-}89.3)^2$	= 132.7
100	$(100\text{-}89.3)^2$	= 114.5
40.476	$(40.476\text{-}89.3)^2$	= 2383
91.304	$(91.304\text{-}89.3)^2$	= 4.016
		3327.15
	$\sum$	3327.15

Fuente: [Elaboración Propia]

Entones VP= $\sqrt{\frac{3327.15}{10}} = 18.24$; $\mu_0 = 85$; $\mu_1 \neq 85$

μ =85; n=10; Promedio muestral PM=89.3%;

Se toma como nivel de confianza del 95%(1- α)

$$Z_c = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{89.3 - 85}{18.24/\sqrt{10}} = 2.3$$

Región Crítica

R.C.] - ω ; $-Z_{0\,975}$ [U] $Z_{0\,975}$; ω [RH_0

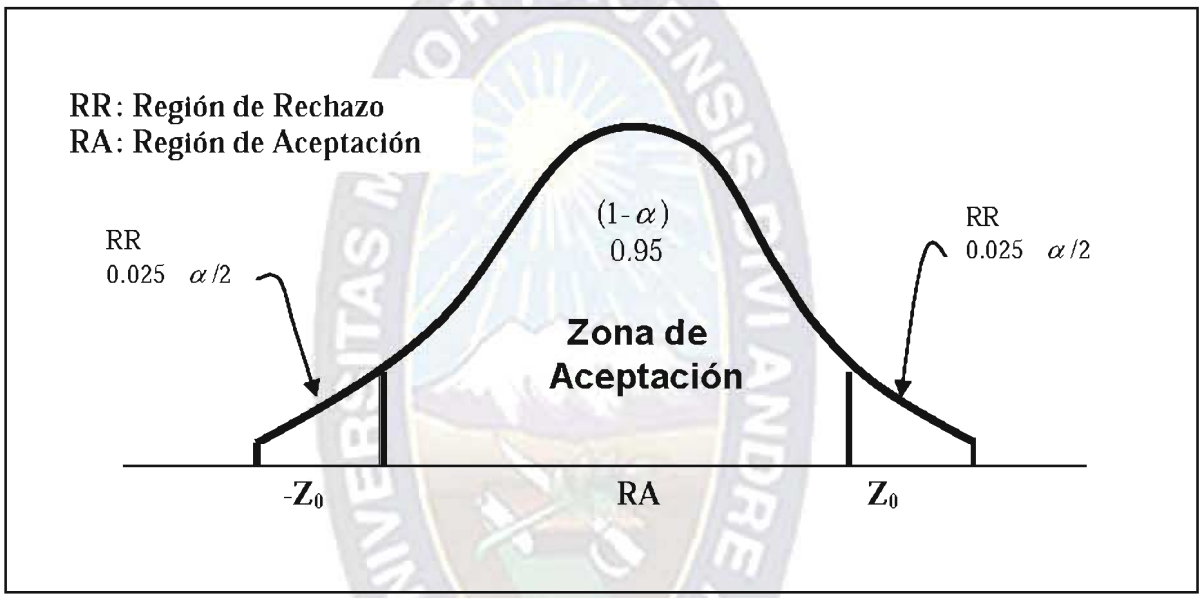
R.C.] - ω ; -1.96 [U] 1.96 ; ω [RH_0

Se puede verificar que el estadístico Z_c=2.3 esta en la región de rechazo, a continuación explicaremos como se tomo esta decisión. El resultado en la región critica se debe verificar

en cual de los intervalos de la región critica se encuentra, para luego analizar este resultados, así rechazamos o aceptamos la hipótesis que nos habíamos planteado, en nuestro caso el $Z_c \in]\ 1.96\ ;\ \omega\ [$ esta en la región de rechazo ver figura 4.11, por lo que se rechaza la hipótesis planteada en este caso la hipótesis nula.

Veamos gráficamente en la Figura 4.14 como se tomo la decisión:

Figura 4.14 Estadístico de toma de decisión



Fuente: [Elaboración Propia]

RH_0 Concluimos que lo estimado no es cierto del 85% de factibilidad.

Observando los resultados del prototipo se tiene resultados de reconocimiento de texto manuscrito continuo en un nivel fiabilidad o reconocimiento del 89.3 por ciento en promedio, considerando las restricciones de la escritura manuscrita planteados, siendo así se puede decir que tenemos alrededor de un 90 por ciento de fiabilidad en el reconocimiento de texto manuscrito en el idioma aymara, con respecto a las tesis con relación al reconocimiento de patrones en otras aplicaciones se tiene resultados aplicados en redes neuronales de un promedio de 88.93 en reconocimiento este resultado esta sujeto también a las restricciones en el reconocimiento de texto que considero el investigador de la tesis.

5 CONCLUSIONES

5.1 CONCLUSIONES

Haciendo uso de los modelos estadísticos: Modelos de Harkov capa oculta (HMMs), Autómatas Estocásticos de Estados Finitos (AEEF), se pudo implementar el reconocimiento de texto manuscrito continuo en el idioma aymara satisfactoriamente, ya que este es el objetivo de la investigación.

Se obtuvieron resultados alentadores de reconocimiento, sobre un buen número de muestras, como los resultados de reconocimiento de las imágenes obtenidas mediante off-line, se realizaron pruebas de corrida del prototipo como en los casos 1 al 10 del capitulo 4.

Obteniendo una muestra n=10 en la prueba del modelos se tiene en promedio alrededor de reconocimiento del 90 por ciento, si solo consideraríamos las restricciones en la obtención de datos esperaríamos resultados en un 100 por ciento. En este tipo de investigaciones por general siempre presenta un nivel de error a un que considere escrituras manuscritas sin restricciones, se puede presentar problemas como se menciono en el capitulo 3 figura 3.3, a la figura 3.7.

Utilizando los modelos de Markov de capa oculta en el reconocimiento de texto manuscrito continuo se llego a probar la hipótesis con resultados mas del 85 por ciento de fiabilidad en reconocimiento de texto.

A través del reconocimiento óptico de caracteres (OCR), en el procesamiento de imágenes off-line, con los modelos planteados se implemento satisfactoriamente.

Se desarrollo el modelo en reconocimiento de texto manuscrito continuo utilizando HMMs, AEEF y N-gramas aplicadas a la lengua aymara.

Se investigo las bases teóricas para el desarrollo de reconocimiento de texto manuscrito continuo (RTM) mediante los modelos HMMs, AEEF y el modelado de lenguaje de N-gramas.

5.2 RECOMENDACIONES

Para desarrollos futuros sobre esta investigación en el reconocimiento de patrones consideramos los siguientes aspectos.

- Considerar el estudio sobre la normalización de alturas de trazo poco comunes o escrituras muy alargadas verticalmente, escrituras manuscritas en diferentes tipos de papel (cuadriculada), para el estudio de niveles de ruido, tratamiento de escrituras manuscritas sobrescritas ya que es un análisis complejo y normalización de escritura con pendientes diferentes izquierda i derecha con correcciones de slant adecuando a la vertical.
- Considerar sistemas de trascripción multimodales que aprovechen la imagen de texto.
- Realizar sistemas borrosos recurrentes mediante estrategias evolutivas en el reconocimiento de patrones y señales, aplicando algoritmos genéticos utilizando Pittsburg y Michigan.
- Investigar sobre los modelos de Markov de procesos de decisión de markov parcialmente observables (POMDPS) con modelos de transición para cada estado.
- Realizar reducciones de probabilidades de error en transmisión en la codificación y decodificaron de vitervi.
- Investigar la relación de procesos de decisión de markov (MDP) con los procesos de decisión de markov parcialmente observable (POMDPS).
- Implementación de algoritmos en otras plataformas, en este trabajo se realizo sobre la plataforma matlab, se puede aumentar la eficiencia computacional acortando los tiempos si aplicamos en implementaciones paralelas.
- Realización de un estudio comparativo de la propuesta de la tesis con redes neuronales u otras tecnologías, estudiando las ventajas y desventajas de ambas aproximaciones al problema.

GLOSARIO DE TERMINOS

- **Inteligencia artificial (IA)**

En una primera aproximación, se puede definir la *inteligencia artificial* como la rama de la computación que estudia la automatización del comportamiento inteligente. La investigación en este campo ha llevado al desarrollo de herramientas computacionales específicas, entre las cuales se cuentan una gran diversidad de formalismos de representación de conocimientos y de algoritmos que los aplican, además de los lenguajes, estructuras de datos y técnicas de programación utilizados para su implementación.

- **Reconocimiento de Formas (RF).**

Estudia como las máquinas pueden observar el entorno, aprendiendo a distinguir patrones de interés en el mismo y posteriormente tomar decisiones razonables acerca de las categorías de esos patrones distinguidos.

- **Algoritmos Genéticos**

Algoritmo capaz de evolucionar hasta llegar a optimizar una respuesta.

- **Algoritmo de Viterbi**

Calcula la trayectoria mas probable en un HMM.

- **Algoritmo forward y backward**

Calculan la probabilidad de una secuencia de palabras.

- **Algoritmo forward – backward (Baum-Welch)**

Estima las probabilidades asociados a un HMM.

- **Entrenamiento**

Proceso en el cual se utiliza HMMs para el reconocimiento de un patrón de características por medios probabilísticas.

- **Estocásticos**

Que tiende a ser probabilística.

- **Interfaz Hombre-Maquina**

Ciencia que estudia la comunicación entre hombre-maquina.

- **Matlab**

Software de simulación para procesos matemáticos

- **Gramática Léxica**

El modelo léxico considera la probabilidad de que cada palabra concreta pertenezca a una categoría gramatical, independientes del contexto que la rodee.

- **Tahuantinsuyu**

El Tawantinsuyö o Cultura *Inka*, desde el punto de vista geográfico, fue el más extenso en relación a las otras culturas nativas de América; abarcó desde *Ankasmayő* (río azul) al sur de Colombia hasta *Maulimayő* (río Mauli) al sur de Santiago de Chile, incluyendo los diferentes pisos ecológicos (costa, sierra y selva) que en la actualidad es territorio de seis países sudamericanos, como son: Perú, Bolivia, Ecuador, parte de Colombia, de Chile y Argentina.

- **Grafemario Unificado**

La lengua de la cultura aymara es ágrafa, lo que no significa que no tuviera escritura, ya que esto es un tema de investigación, sin embargo es un idioma propio que por convención ha utilizado la simbología latina para su escritura.

- **Grafemario Aymara**

EL aymara tiene 26 consonantes y 3 vocales. Además cada vocal tiene la posibilidad de un alargamiento vocálico ä(aa) y ü(uu).

El termino **grafemario** unificado recoge las normalizaciones internacionales de la disciplina lingüística. En esta disciplina se define la unidad mínima de la escritura de grafema es por eso que se prefiere hablar de un grafemario y no de un alfabeto.

- **Fonemas.**

Propio del aparato sonorizado del ser humano.

- **Alófonos**

Las vocales e, o en la escritura aymara no están definidas, siendo la pronunciación de una determinada frase utilizando estas mismas vocales.

- **Lenguaje de n-Gramas**

Mecanismo para definir la estructura del lenguaje, es decir, para restringir adecuadamente las secuencias de unidades lingüísticas más probables.

- **Población**

Población en lengua aymará según departamento: La Paz 975.440, Oruro 106.030, Potosí 55.893 y Cochabamba 58.055 según instituto nacional de estadística 2005 (INE).

- **Lenguaje**

Lenguaje Don del hombre de expresar por medio de sonidos articulados lo que piensa o siente

- **Idioma**

Idioma comprende toda la estructura del habla como de la escritura

REFERENCIA BIBLIOGRÁFICA

- [Hernández S.; Roberto F.; Fernández C.;2006]. Hernández Sampieri, Roberto. Fernández Collao Carlos. Baptista Pilar Lucio. Metodología de la Investigación 4ta Edición. México. Infagon Web, 2006. 849_p
- [Quisbert M.; 1190]Quisbert, Marcelo. Quechua y aymara. 1º Edición. La Paz Bolivia ,1990. 50_p
- [Reynaga B.; 2005]. Reynaga Burgoa Ramiro. Tawaintisuyu. 6ta Edición. La paz Bolivia. Producciones graficas culturales, 2005. 369_p
- Oncina, José. Aprendizaje Computacional y Extracción de Información. 2007, 47_p
- [Jelinek F.; 1997] Jelinek, F. Métodos Estadísticos para el reconocimiento del habla. Prensa de MIT. 1997, 40_p
- [Peinado V.; 2004]. Peinado, Víctor. Modelos de Lenguaje Estadísticos.2004
- [Bergasa P.; 1998]Bergasa Pascual, Luís Miguel. Introducción a los Modelos Ocultos de Markov. 1998. 50_p
- [Rodríguez F.; 2006]. Rodríguez, F., Bautista S. Modelos Ocultos de Markov Para el Análisis de Patrones Especiales. 2006.
- [Rabiner L.; 2005] Rabiner, L. Programa de Entrenamiento Sobre Modelos de Markov Escondidos y Aplicaciones Seleccionados en el Reconocimiento de Habla. Actas del IEEE, 77(2). 257-286.
- [Fine S.; 1998]. Fine S, Singer Y.EL Modelo de Hidden Markov Jerarquico. 1998. 41-64.
- [Doménech J.; 2000]. Doménech, Javier. Estudio de Técnicas de Preproceso para el reconocimiento de Texto Manuscrito. Universidad Politécnica de Valencia. 2000.